Presentation matters: Evaluating speaker identification tasks

Benjamin O'Brien, Christine Meunier, Alain Ghio

Presentation matters: Evaluating speaker identification tasks

Benjamin O'Brien[1], *Christine Meunier*[1], *Alain Ghio*[1]

[1]Aix-Marseille Univ, CNRS, LPL, UMR 7309

benjamin.o-brien@univ-amu.fr

Abstract

This paper details our evaluations and comparisons of speaker identification (SID) performance by listeners across different tasks. Experiment 1 participants completed traditional target-lineup (1-out-of-N speakers or out-of-set speaker) and binary (speaker verification) tasks. Experiment 2 participants completed trials online by using a *clustering* method by grouping speech recordings into speaker-specific clusters. Both studies employed similar speech recordings from the PTSVOX corpus. Our results showed participants who completed the binary and clustering tasks had higher accuracy than those who completed the target-lineup task. We also observed that independent of the tasks participants found some speakers significantly more difficult to identify relative to their foils. Pearson correlation procedures showed significant negative correlations between accuracy and task-dependent temporal-based metrics across tasks, where an increase in time required to make determinations yielded a decrease in perceptual SID performance. These findings underscored the important role of SID task design and the process of selecting speech recordings. Future work aims to examine the relationship between different perceptual SID task performances and scores generated by automatic speaker verification systems.

Index Terms: speaker identification, presentation methods, accuracy and temporal correlates

1. Introduction

There are a number of well-documented challenges that listeners face when tasked to identify speakers [1] [2] [3]. In addition to developing corpora of high-quality speech recordings that capture idiosyncratic speaker characteristics, the methods used to present speech materials are also influential on listeners [4] [5].

The collection, selection, and presentation of speech materials is at the core of research pursued by forensic phoneticians [6]. Decisions regarding which speech recordings to select and their sequence have the potential to influence witnesses and interpretations of evidence. Some examples that have been studied include accent [7] and prosody [8]. With respect to presenting materials and findings at judiciary proceedings, it is of upmost importance to eliminate or minimise any biasing features.

Oftentimes forensic phoneticians develop speech recording lineups, as a way of collecting evidence. However, the process of transforming a visual (facial) lineup into an auditory (speech) one is difficult. Major concerns have been addressed over the design of "voice parades" or "earwitness lineups," where witnesses are presented a collection of speech recordings and are tasked to identify whether a target is present in the speaker set (1-out-of-N) or absent (out-of-set). Numerous studies have detailed that listeners were inefficient at identifying targets [9], and that the use of speech lineups yielded less-accurate and consistent results based on lineup constraints [10, 11]. Factors that contribute to these loses include, among others things, the process of selecting (consistent) phonetic content for target and foil speakers and speech recording quality. The lack of a consensus surrounding the number of voices to include in a lineup is problematic.

A much simpler perceptual SID method employs a binary approach, where listeners are tasked to determine whether two speech recordings belong to the same speaker or not. This approach is the basis of automatic speaker verification system models. Previous perceptual SID studies have used this method to examine the effects of such things as noise [10], language familiarity [12] [13], and stimuli selection methods [14]. Oftentimes this approach requires numerous tests, which can be time-consuming for listeners. In addition, there are concerns that binary-based perceptual SID tasks are encumbered by memory bias, as listeners may perceive or recall speech characteristics relating to speech recordings in previous binary trials.

As an alternative to these approaches, we proposed the development of a perceptual *clustering* method, which is often employed in the domain of machine learning [15] [16]. In general listeners are tasked to cluster similar speech recordings into speaker-specific clusters. Promising findings from an earlier study were reported [17], as participants were able to personalise their engagements with speech materials and organise their proximities in relation to their perceived likeness. The task was designed to be open and more evocative of natural interactions between speech listeners, as opposed to more artificial tasks that restrict listener expression.

The goal of this study was to examine the effects of perceptual SID task design on performance by listeners. It was of interest to study whether design features, such as the method of response and the number of stimuli presented per trial, influenced SID accuracy. Our second goal was to study the relationship between perceptual SID performance (accuracy) and task-dependent temporal-based metrics. While we hypothesised that there was a direct relationship between the number of presented stimuli and the time required to make informed discriminations based on speech materials, we wanted to study whether they correlated to accuracy.

2. Methods

2.1. Stimuli

Speech recordings from 10 female and 10 male native-French speakers were selected from the PTSVox database [18]. The age range of speakers was 18 to 24 years (mean age 19.7 $\pm$ 1.6 years). They recited three French-texts, which were recorded with a Zoom H4N stereo microphone (sampling rate: 44.1 kHz; bit depth: 16-bit).

For each speaker we extracted 24 speech fragments (*utterances*) from the speech recordings using the speech-analysis software Praat [19]. The duration of the extractions ranged from 1.062 to 3.536 s (mean duration 1.875 $\pm$ 0.379 s). All 480

speech recordings were normalised to 0 dB by using a custom script written in MATLAB 2016b (MathWorks Inc, USA).

Five female and five male speakers were randomly assigned to an inset speaker group, while the remaining speakers were assigned to an out-of-set group. Similarly, twelve utterances were evenly distributed to the inset and out-of-set speaker groups. These speech recording sets were used across the experiments.

2.2. Participants

35 native-French speakers (27 female) participated in Experiment 1 (mean age 26.2 $\pm$ 8.0 years). A different group of 19 native-French speakers (17 female) participated in Experiment 2 (mean age 26.2 $\pm$ 6.4 years). All participants reported good hearing. All participants consented to voluntary participation in the study and were compensated for their time.

2.3. Procedure

2.3.1. Experiment 1

Experiment 1 participants completed two tasks, which are described below. For both tasks participants completed a series of trials programmed in Lancelot / Perceval [20] on desktop computers at CEP-LPL. The gender of the speech recordings alternated for the two tasks, which were counter-balanced between participants. Throughout the study participants wore AKG K702 headphones. Prior to testing, participants listened to a speech recording and adjusted the volume to their comfort.

Participants completed 30 Target-Lineup (TL) task trials. For each trial, they were presented a Target speech recording and five speech recordings that constituted a Lineup. The Target speaker utterance was different from the Lineup speakers, who all spoke the same utterance. Participants were tasked to determine whether the Target speaker was present in the Lineup. To make these determinations, they used a custom interface (Fig. 1). If they believed the Target was absent from the Lineup, participants selected the item below a red 'X'. Participants had the option of using a row of boxes positioned above the Lineup speakers to aid their selection process. Participants were unlimited by the number of speech recording listens.

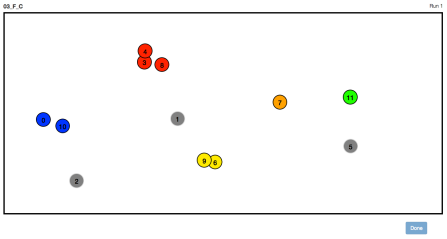
Figure 1: *Target-Lineup trial interface*

Participants completed 100 Same-Different (SD) task trials. Each SD trial began with a short sound (0.8 s) produced by a sinusoidal oscillator with a frequency of 500 Hz. Following 2 s of silence, a speech recording was automated and synchronised with *Voix A* text displayed in a yellow rectangle on the computer screen. The recording was proceeded by 0.6 s of silence with a black screen (no image) followed by a second speech recording with *Voix B* in a blue rectangle. The gender of the two speakers was the same, while the utterances were different. Participants then had 5 s to determine whether the two speech recordings belonged to the same speaker or two different speakers by pressing the right or left button, respectively.

2.3.2. Experiment 2

Experiment 2 participants completed a series of 10 Cluster trials on their personal computers by accessing a website that provided them with a custom interface developed at Laboratoire Informatique d'Avignon, Université du Vaucluse-Avignon (Fig. 2). Each trial was composed of 12 speech recordings, which were represented as numbered circles. Participants were tasked to listen to each recording (unlimited) and classify it into a cluster, which represented a unique speaker. To classify the speech recordings, participants were instructed to right-click on the circle, which revealed a drop-down menu with different classification colors. The minimum and maximum number of clusters permitted per trial were two and six, respectively. Participants were encouraged to use personal headphones and were provided detailed instructions on how to complete the task and use the interface.

Figure 2: *Cluster trial interface*



2.3.3. Trial design

For the TL task trials, each inset speaker (5) was presented as the Target six times. For half of these trials, the inset speaker randomly replaced an out-of-set speaker in the Lineup. For each inset speaker trial (6), six inset and six out-of-set utterances were randomly selected (non-repeating). Over the 30 TL task trials, all inset and out-of-set utterances were repeated five times (balanced).

The speech recordings selected for the SD and Cluster task trials were based on those used in the TL task trials. For the SD task trials, 12 trials included each inset speaker (60 of the 100 total). For each out-of-set speaker (5), we identified two TL trials when it was compared against the selected inset speaker and counter-balanced their juxtaposition (10 trials). For the remaining two trials, we randomly selected two TL trials when the selected inset speaker was included in the Lineup during the TL trials. As it was important to present materials with 1:1 ratio same-to-different, the remaining SD task trials (40 of the 100 total) featured trials where an out-of-set speaker was randomly selected and recited two random (non-repeating) utterances.

For the Cluster task trials, each inset speaker was included in two trials (10 total). The six TL trials when the inset speaker served as the Target were randomly divided across these two trials. One out-of-set speaker was randomly selected to be included in both trials, whereas the remaining four out-of-set speakers were randomly divided into two groups and assigned to a trial. Per trial each out-of-set speaker (3) was randomly assigned between 2 and 5 Distractor utterances from the TL task trials. The speech recordings were unique across the two Clus-

ter trials per inset speaker.

2.4. Data processing

A *trial score* metric was used to measure SID performance across tasks. For both TL and SD task trials participants received a *1.0* score for *true* responses and a *0.0* for all *false* responses. For Cluster task trials, if a cluster contained an inset speaker speech recording, then the cluster received a ratio score based on the number of inset speaker speech recordings in the cluster divided by the total number of speech recordings in the cluster.

In addition we measured different task-dependent temporal-based metrics to examine whether they correlated to performance using Pearson correlation procedures. For TL trials we calculated the trial duration (s). For the SD task trials we calculated the reaction time (s). For the Cluster trials we calculated the average number of listens per cluster with an inset speaker.

Because Experiment 1 participants performed both TL and SD task trials, separate ANOVA procedures were applied based on speaker stimuli gender. Mixed ANOVA procedures were carried out with the task type (Target-Lineup, Same-Different, Cluster) as the between-subject factor and inset speaker as the inter-subject factor (α = 0.05). Where main effects were detected, post-hoc Bonferroni-adjusted t-tests were carried out.

2.5. Preliminary analysis

Normal distribution functions were fitted to the average duration to complete the TL and Cluster task trials, while reaction time was used to evaluate participant normalcy during the SD trials. All participant data was included in the study with the exception of data collected from two participants during the SD task trials, as their means were greater than three standard deviations from their group means. Table 1 illustrates these first findings.

Table 1: *Mean duration across task types*

Task type	Mean (s)	SD (s)	Stimuli per trial (#)
Target-Lineup	27.674	8.602	6
Same-Different	1.29	0.271	2
Cluster	201.882	69.273	12

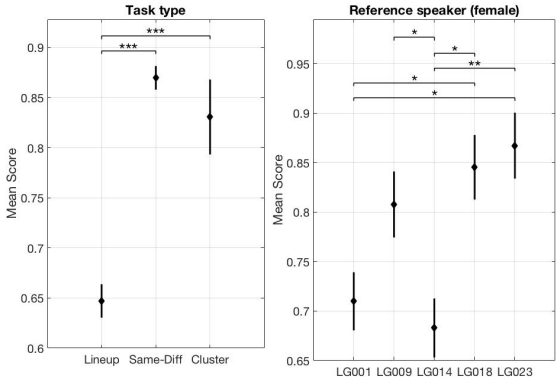
In general we observed similar trends between the SD and Cluster tasks, where the sum of true positive (44%) and negative (44%) responses for the SD task was similar to the true positive responses (86%) for the Cluster tasks. In addition both had similar false positive and negatives responses (6-7%). In comparison, the TL task had a similar true positive response (43%), but we observed a decrease in true negative (25%) and false negative (2%) responses and an increase in false positive responses (30%).

3. Results

3.1. Task and speaker comparisons

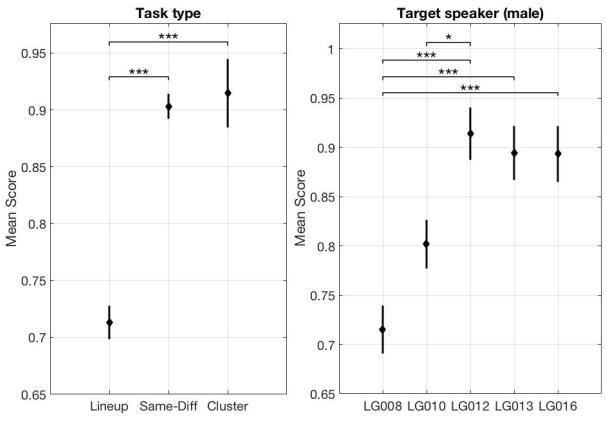
For female speaker stimuli we found main effects for mean score on task type $F_{2,1619}$ = 59.58, p $<$ 0.001, η_p^2 = 0.07 and speaker $F_{4,195}$ = 6.93, p $<$ 0.001, η_p^2 = 0.02, but no interactions, p $>$ 0.05. Post-hoc t-tests revealed that participants had significantly greater accuracy when performing the SD (0.87 $\pm$ 0.01) and Cluster (0.83 $\pm$ 0.04) tasks when compared to the TL task (0.65 $\pm$ 0.02), p $<$ 0.001 (Fig. 3-Left). Fig. 3-Right illustrates the significant differences between female speakers across tasks.

Figure 3: *ANOVA results with female speaker stimuli: Tasks (Left) and Target speakers (Right). Diamonds and vertical lines represent means and standard errors, respectively.* {*, **, ***} *represent p* $<$ {*0.05, 0.01, 0.001*}



We found main effects on male speaker stimuli for mean score on task type $F_{2,1617}$ = 56.75, p $<$ 0.001, η_p^2 = 0.07 and speaker $F_{4,1617}$ = 10.93, p $<$ 0.001, η_p^2 = 0.03, but no interactions, p $>$ 0.05. Post-hoc t-tests revealed that participants had significantly greater accuracy when performing the SD (0.9 $\pm$ 0.01) and Cluster (0.91 $\pm$ 0.03) tasks when compared to the TL task (0.71 $\pm$ 0.01), p $<$ 0.001 (Fig. 4-Left). Fig. 4-Right illustrates the significant differences between male speakers across tasks.

Figure 4: *ANOVA results with male speaker stimuli: Tasks (Left) and Target speakers (Right). Diamonds and vertical lines represent means and standard errors, respectively.* {*, **, ***} *represent p* $<$ {*0.05, 0.01, 0.001*}



3.2. Temporal correlates

Based on participant responses we calculated the mean score and trial duration (TL), response time (SD), and number of listen (Cluster) means for each inset speaker. Pearson correlation

procedures were then applied to these metrics. Table 2 illustrates these findings, where {*, **, ***} represent p $<$ {0.05, 0.01, 0.001} significance.

Table 2: *Pearson correlations between score and task-dependent temporal-based metrics*

	Target-Lineup		**Same-Different**		**Cluster**	
Stimuli	p	ρ	p	ρ	p	ρ
Female	***	-0.79	**	-0.34	-0.27	**
Male	***	-0.84	***	-0.46	-0.17	*

4. Discussion

The goal of our study was to examine whether the design of perceptual SID tasks influenced the performance of listeners. Our findings showed participants performed with higher accuracy when they completed the SD and Cluster tasks rather than the TL task. The TL task was distinguished from the other tasks due to the possibility that the Target speaker was absent from the Lineup, which contributed to the increase in false positives. It was clear that when they were presented the out-of-set option, participants exhibited a tendency to make inset selections.

While speech recordings and comparisons were consistent across tasks, we reported that participants found some inset speakers more difficult to identify with respect to their foils. When we compared score means per task between genders, we observed the following: TL task (female: 65%, male: 71%); SD task (female: 89%, male: 87%); and Cluster task (female: 82%, male: 90%). While SD task participants performed slightly better when they were presented female speech recordings, we observed the opposite for TL and Cluster participants, who improved performance when presented male speech recordings. We reported no significant interactions between task and speakers, which suggest the tasks did not influence performance of certain inset speakers (see: female speakers LG001 and LG014 and male speakers LG008 and LG010). Future work aims to gain a better understanding of these differences by measuring the acoustic similarities between target and foil speech recordings. By doing so we might gain insight into acoustic-perceptual correlates in relation to SID task designs.

Our application of Pearson correlation procedures to perceptual SID performance (accuracy and task-dependent temporal-based metrics) revealed similar trends across tasks, where we reported negative correlations across all tasks. This finding supports traditional hypotheses that posit that when listeners perceive stimuli as similar, they require more time to make determinations, which can affect (and reduce) their accuracy. The TL task boasted the strongest ρ-values. Unlike the SD task, where the window to respond was limited to a 5 s, TL task participants were allowed unlimited time to (re-)listen and make their determinations. However, our findings showed that this flexibility in fact was detrimental to SID performance, which, in the context of forensics, brings into question the reliability of target-lineup procedures, especially when considering the subjective limitations of listener retention.

Interestingly, although the Cluster task reported high score means, we observed near-flat correlations for both female and male speech recordings, which suggest that the task is quite variable and dependent on the listeners and their perceptual capacities and limitations. A major reason for developing the Cluster task was because it was not restrictive like the other tasks and it allowed listeners to engage with the speech materials freely. Thus this openness lent itself to participants employed different listening strategies, which limited range of mean listens per cluster. To understand the Cluster task and its potential, we plan future work to examine the mixed-effects of stimuli selection, number of stimuli per trial, limited number of listens, and trial duration.

To expand upon these findings our future work aims to integrate scores generated by using automatic speaker verification (ASV) systems. In general, ASV systems can be used to compare speech recordings and produce scores that indicate the likelihood that they were produced by the same speaker. Several studies have been developed to examine the relationship between subjective (human) and objective (machine) SID performance. A major work by [21] was developed to examine how humans can effectively use ASV systems and their potential in the domain of forensics. A study by [22] showed perceptual SID performance correlated strongly to ASV models trained with phonetic features (F0, F1-F4). In a previous study [17] we reported significant correlations between cosine distance scores generated from a custom ASV system based on i-vectors and the accuracy of participants tasked to cluster speech materials. Since then we have developed an ASV model that relies on x-vector speaker embeddings and probabilistic linear discriminant analysis [23] to produce log-likelihood ratio values between corresponding x-vectors. We have trained this model with two separate corpora and our results are forthcoming. The goal is to examine whether the perceptual SID task plays a role in the strength (and significance) of subjective-objective correlations.

5. Conclusions

This paper detailed our development of three perceptual SID tasks with similar speech recordings. Our findings underscored the effect of SID task design on accuracy, as well as its relationship to task-dependent temporal metrics. Although optimising human and machine SID performance is important, its value depends on whether necessary constraints provide the contexts that allow correct determinations to be made. Our findings add to developing research that focuses on the perceptual capacities of listeners and how they might be used to model ASV systems. From a different perspective, ASV systems might be modelled differently to optimise listener SID performance given these perceptual limitations. These findings suggest their applications in ASV-design modeling. For example, as we reported that the target absent option led to an increase in false positive responses, which, in turn, significantly decreased accuracy, future research might examine the threshold of this negative effect with respect to the number of speakers in a set. As we reported that listeners performed the binary and clustering tasks quite well, it might be of interest to study their performance when they received feedback based on knowledge of performance. Examining these changes in relation to our findings might provide further insight on the effects of perceptual SID task design and their use in modelling ASV systems.

6. Acknowledgements

This work was funded by the French National Research Agency (ANR) under the VoxCrim project (ANR-17-CE39-0016). The authors thank the CEP staff (www.lpl-aix.fr/ cep), especially, Carine André, for assisting in the perceptual experiments.

7. References

[1] Cambier-Langeveld, Rossum, and Vermeulen, *Whose voice is that? Challenges in forensic phonetics*, 01 2014, pp. 14–27.

[2] Mattys, Davis, Bradlow, and Scott, "Speech recognition in adverse conditions: A review," *Language and Cognitive Processes*, vol. 12, no. 7-8, pp. 953–978, 2012.

[3] F. Nolan, "Speaker identification evidence: its forms, limitations, and roles." Law and Language: Prospect and Retrospect, December 2001.

[4] L.-J. Boe and J.-F. Bonastre, "L'identification du locuteur: 20 ans de témoignage dans les cours de justice le cas du lipsadon "laboratoire indépendant de police scientifique"," in *JEP-TALN-RECITAL*, vol. 1, 06 2012, pp. 417–424.

[5] H. Hollien, R. Bahr, H. Kunzel, and P. Hollien, "Criteria for earwitness lineups," *International Journal of Speech Language and the Law*, vol. 2, pp. 143–153, 04 2013.

[6] J. Olsson, *What is forensic linguistics?* England: Nebraska Wesleyan University., 2003.

[7] M. Sloos, A. Ariza García, A. Andersson, and M. Neijmeijer, "Accent-induced bias in linguistic transcriptions," *Language Sciences*, vol. 76, p. 101176, 2019, biases in Linguistics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0388000117303467

[8] M. Harris, S. Gries, and V. Miglio, "Prosody and its application to forensic linguistics," *Linguistic Evidence in Security, Law and Intelligence*, vol. 2, 12 2014.

[9] H. Hollien, "Perceptual identification of voices under normal, stress, and disguised speaking conditions," *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, vol. 56, 11 1974.

[10] H. M. J. Smith, K. Bird, J. Roeser, J. Robson, N. Braber, D. Wright, and P. C. Stacey, "Voice parade procedures: optimising witness performance," *Memory*, vol. 28, no. 1, pp. 2–17, 01 2020. [Online]. Available: https://doi.org/10.1080/09658211.2019.1673427

[11] J. Mullennix, A. Ross, C. Smith, K. Kuykendall, J. Conard, and S. Barb, "Typicality effects on memory for voice: Implications for earwitness testimony," *Applied Cognitive Psychology*, vol. 25, pp. 29–34, 01 2011.

[12] D. Fleming, B. Giordano, R. Caldara, and P. Belin, "A language-familiarity effect for speaker discrimination without comprehension," *Proceedings of the National Academy of Sciences*, vol. 111, 09 2014.

[13] S. Levi and R. G. Schwartz, "The development of language-specific and language-independent talker processing," *Journal of speech, language, and hearing research : JSLHR*, 06 2013.

[14] C. Mühl, O. Sheil, L. Jarutyte, and P. Bestelmeyer, "The bangor voice matching test: A standardized test for the assessment of voice perception ability," *Behavior Research Methods*, vol. 50, pp. 1–9, 11 2017.

[15] T. Kinnunen and T. Kilpeläinen, "Comparison of clustering algorithms in speaker identification," *Proc. IASTED Int. Conf. Signal Processing and Communications (SPC)*, pp. 222–227, 01 2000.

[16] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," 09 2016, pp. 1–6.

[17] B. O'Brien, A. Ghio, C. Fredouille, J.-F. Bonastre, and C. Meunier, "Discriminating speakers using perceptual clustering interface," in *Proc. XVII AISV Conference: Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*, 2021.

[18] A. Chanclu, L. Georgeton, C. Fredouille, and J.-F. Bonastre, "Ptsvox: une base de données pour la comparaison de voix dans le cadre judiciaire," in *6e conférence conjointe Journées d'Études sur la Parole*, 2020, pp. 73–81.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," vol. 5, no. 9/10, pp. 341–345, 2001.

[20] C. Andre, A. Ghio, C. Cavé, and B. Teston, "Perceval: a computer-driven system for experimentation on auditory and visual perception," in *International Congress of Phonetic Sciences*, 06 2007, pp. 1421–1424.

[21] C. S. Greenberg, A. Martin, L. Brandschain, J. Campbell, C. Cieri, G. Doddington, and J. Godfrey, "Human assisted speaker recognition in nist sre10," in *Odyssey*, 2010.

[22] F. Kelly, A. Alexander, O. Forth, S. Kent, J. Lindh, and J. Åkesson, "Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features," in *INTERSPEECH*, 2016.

[23] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.