Discriminating speakers using perceptual clustering interface

Benjamin O'Brien, Christine Meunier, Alain Ghio, Corinne Fredouille, Jean-François Bonastre, Carolanne Guarino

BENJAMIN O'BRIEN, CHRISTINE MEUNIER, ALAIN GHIO, CORINNE FREDOUILLE, JEAN-FRANÇOIS BONASTRE, CAROLANNE GUARINO

Discriminating speakers using perceptual clustering interface

The challenges facing naïve listeners tasked with identifying or discriminating speakers are well documented. In addition to providing listeners with high-quality speech recordings that accurately represent the speakers, the perceptual task itself is equally important. Conventional perceptual speaker identification and discrimination tasks include voice parades and pairwise comparisons, however, there are concerns regarding their design and memory biases, respectively. As an alternative our study proposed the development and use of a perceptual *clustering* method, where participants performed speaker discrimination tasks with a novel clustering interface. A state-of-the-art automatic speaker verification (ASV) system was used to select speech stimuli used in our study. Our findings revealed participants were able to distinguish speakers with high accuracy, which significantly correlated with scores generated by our ASV system.

Keywords: Speaker identification, Naïve listeners, Automatic speaker verification systems, Clustering.

1. *Introduction.*

When naïve listeners are tasked with identifying speakers, they first compare voice characteristics by discriminating speech sounds and then make judgements as to whether they are similar or different. This perceptual process of identifying voices as similar (Gerlach, McDougall, Kelly, Alexander, and Nolan, 2020) is part of a larger vein of research on voice perception (Belin, Bestelmeyer, Latinus, and Watson, 2011). Substantial research has been devoted to identifying which acoustic and phonetic features are used for voice perception. LaRiviere (1971) demonstrated that the fundamental frequency (F0) and second and third formants all play important roles in speaker identification. Baumann and Belin (2008) reported similar evidence of the effect of F0 on speaker identification judgements. Roebuck and Wilding (1993) showed listeners improved speaker identification judgements when they were presented speech sounds that contained more vowels in comparison to those with more consonants. Lindh and Eriksson (2010) also reported the effect of speech tempo on speaker identification evaluations. However, there remain many questions regarding voice perception and the effects of such things as speech modalities (Blatchford and Foulkes, 2006; Hollien, Majewski, and Doherty, 1982) or environment (Smith, Baguley, Robson, Dunn, and Stacey, 2018;

Kerstholt, Jansen, Amelsvoort, and Broeders, 2006). Recent research reviews by Stevenage (2018), Mattys, Davis, Bradlow, and Scott (2012), and Kreiman and Van Lancker Siditis (2011) demonstrate the growing interests to improve our understanding of voice perception and its applications in neuroscience, forensics, linguistics, and computer science domains.

Another obstacle to consider is the situation in which naive listeners are tasked with identifying speakers. Retention period (Boë and Bonastre, 2012; Hollien, Bahr, Künzel, and Hollien, 2013), speech qualities (Sloos, García, Andersson, and Neijmeijer, 2019; Harris, Gries and Miglio, 2014) or environmental factors (Olsson, 2003) are all important contextual factors that have been shown to affect speaker identification performance. To study their effects on speaker identification performance, various perceptual tasks have been developed. A distinguishing characteristic of a perceptual speaker identification task centres on whether listeners are tasked with identifying a target speaker or discriminating between speakers. While identification involves the process of assessing speech similarities between voices, discrimination involves the process of discerning their differences.

With regards to the former, a popular perceptual speaker identification task is a voice parade (Jong, Nolan, McDougall, and Hudson, 2015), where listeners are tasked with determining whether a speech recording in a set belongs to a target speaker. The method transforms the typical facial-recognition set from the visual to the auditory domain. Voice parades have been used in a variety of contexts ranging from psychoacoustics to forensics (Smith, Bird, Roeser, Robson, Braber, Wright, and Stacey, 2020; Kreiman and Van Lancker Siditis, 2011; Mullennix, Ross, Smith, Kuykendall, Conard, and Barb, 2011). An important distinguishing feature of the voice parade method is whether a target-speaker is present in the speaker set or absent (Öhman, Eriksson & Granhag, 2010). While there have been some criticisms of the approach (Hollien et al., 2013), guidelines have been presented in the context of forensic linguistics (Broeders, and Amelsvoort, 1999).

A much simpler perceptual task employs a pairwise comparison method, where listeners make discriminations as to whether two speech recordings are similar or different. The popularity of this approach lies in its simplicity. This task has been used to examine various effects, such as noise (Smith et al., 2018), language familiarity (Fleming, Giordano, Caldara, and Belin, 2014), and speaker familiarity (Baumann and Belin, 2008). One drawback of this method is that it requires numerous tests, which can be time-consuming for listeners, however, Mühl, Sheil, Jarutyte, Bestelmeyer (2017) proposed a "same-to-different" task with approximate 10 min duration. While fatigue can play an influential role on speaker identification performance, so too can memory bias, as oftentimes speech recordings uttered by similar speakers are introduced and re-presented. Moreover, an important study by Jenson and Saltuklaroglu (2021) reported that brain activations made by naïve listeners differed when they were presented same or different speech recording pairs. This observation suggests that the task requires different auditory and decision-making processes depending on

the presented stimuli. It was of interest to develop a perceptual task that treated all judgements about stimuli equally.

Rather than adapting existing perceptual speaker identification tasks that restricted listeners to binary responses, we wanted to design a method that afforded responses that reflected their natural engagements with speakers. A different type of perceptual task is a voice sorting method, where listeners freely organise speech recordings into groups that represent specific speakers. Voice sorting has been used to examine the behavioural responses of familiar and unfamiliar listeners (Lavan, Kreitewolf, Obleser, and McGettigan, 2021; Lavan, Burston, and Garrido, 2018; Stevenage, Symons, Fletcher, Coen, 2019). We used the voice sorting method as a model for our development of a perceptual *clustering* method.

It was believed that a perceptual clustering method would provide an ideal platform for naïve listeners to personalise their engagements with speech recordings. Clustering is a common subject of study in the domain of machine learning (Kinnunen and Kilpeläinen, 2000; Lukic, Vogt, Dürr, and Stadelmann, 2016). In general, clustering algorithms rely on automatic speaker verification (ASV) systems to extract acoustic and phonetic information from speech recordings in order to create speaker models that are then used to produce scores based on evaluations as to how likely two speech recordings belong to the same speaker. It was of interest to use a similar method to examine whether naive listeners performed similarly to scores produced by a state-of-the-art ASV system. Moreover, this method was more economical in terms of the number of trials and repetitions of speech recordings belonging to speakers, and thus minimised the potential for memory bias effects.

The goal of our current study was to understand better voice perception in the context of how naïve listeners discriminate speech materials and make similarity judgements. Our primary goal was to examine whether naïve listeners could effectively use a perceptual clustering method. It was believed that a clustering method provided listeners with a more natural manner in which to engage with speech materials, and any key findings would support its continued use as an alternative to other perceptual speaker identification tasks. Our second goal was to compare human speaker identification performance with scores produced by a state-of-the-art ASV system. It was of interest to examine the relationships between these performances with key implications for how ASV systems might be modelled differently to better reflect human responses to speech stimuli.

2. *Method.*

2.1 Stimuli

As part of the *VoxCrim* project (Chanclu, Georgeton, Fredouille, and Bonastre, 2020), speech recordings were selected from the PTSVox database. Among other recordings, the corpus includes 24 francophone speakers (12 female and 12 male) who all recited three French-texts into a Zoom H4N stereo

microphone (sampling rate: 44.1 kHz; bit depth: 16-bit) over the course of two recording sessions.

In order to select speakers, we applied a popular method in the field of ASV systems, which involves the extraction of acoustic information and compresses these features into i-vectors (Dehak, Kenny, Dehak, Dumouchel, and Ouellet, 2011; Kanagasundaram, Vogt, Dean, Sridharan, and Mason, 2011). These i-vectors can then be used to train and test models, where distance scores calculate a degree of similarity between speech recordings. For our development, we first used the SPro toolkit (Gravier, 2021) to extract and normalise 19 MFCCs, deltas, and delta-deltas (57 total features) from each recording. Using the ALIZE system (Bonastre, Wils, and Meignier, 2005), we compressed these features into speaker i-vectors, where the Universal Background Model was composed of 512 Gaussian Mixed-Model components and the i-vector dimensionality was set to 400. As recent work has shown that Cosine Distance Scoring (CDS) with Within-Class covariance normalisation (WCCM) is effective and accurate at identifying speakers, while reducing the complexity of the task (Fredouille and Charlet, 2014), we calculated the CDS between each i-vector and then the WCCM was computed over the entire set.

Following these operations, a simple Python script was written to select the speaker composition of each group: the *Alpha* group was composed of five speakers with the greatest distance between them, whereas as the *Betha* group was composed of five speakers with the smallest distance between them. Tab. 1 offers relevant information pertaining to the selected speakers, including background information and group assignments. For each speaker, we randomly selected 12 utterances (see Tab. 2) extracted from the speech recordings with a duration range from 1.062 to 3.536 s (120 recordings total with a mean duration 1.47 ± 0.51 s). Each group was divided into three sessions, such that each session was balanced and composed of four different (non-repeating) utterances per speaker.

Table 1 – *Speaker stimuli description*

Speaker	Group	Gender	Age	Region	Smoker?
LG006	Alpha	Female	24	Paris	Yes
LG008	Alpha	Male	24	Lorraine	No
LG018	Alpha	Female	19	Picardie	No
LG023	Alpha	Female	19	Haut-Rhin	No
LG024	Alpha	Male	19	Rhône	No
LG005	Betha	Male	18	Rhône	No
LG013	Betha	Male	22	Loire	No
LG017	Betha	Male	20	Rhône	Yes
LG019	Betha	Male	20	Bourgogne	No
LG021	Betha	Male	20	Auvergne	No

Table 2 – *Utterances*

on trouve une espèce de chat
ils sont noirs avec deux tâches blanches sur le dos
leur poil est beau et doux
vit une colonie d'oiseaux
laisser tomber leurs œufs
ma sœur n'a qu'à traverser la rue
pour rencontrer ces deux espèces
au cœur d'un parc naturel
sur le coup de midi
pour aller observer ces animaux
ma sœur est venue chez moi hier
elle me parlait de ses vacances en mer du Nord
dans notre dos tombé un petit oiseau
ses deux ailes étaient blessées
son cœur battait très vite
son plumage était beau et doux
pour regarder dans la rue
je m'approchais du bord de la fenêtre
il avait dû faire fuir l'oiseau
s'éloignant d'un nid perché sur un arbre
la bise et le soleil se disputaient
quand ils ont vu un voyageur qui s'avançait
faire ôter son manteau au voyageur
serait regardé comme le plus fort
serrait son manteau autour de lui
le soleil a commencé à briller
le voyageur réchauffé a ôté son manteau

2.2 Participants

Twenty-four native-French speakers (14 female and 10 male; ages 24.2 ± 6.7 years) participated in the study. All participants reported healthy hearing and good or corrected vision. All participants consented to voluntary participation in the study and were informed of their right to withdraw at any time. They were compensated for their time. This study was performed in accordance with the ethical standards of the Declaration of Helsinki (Salako, 2006).

2.3 Experimental setup

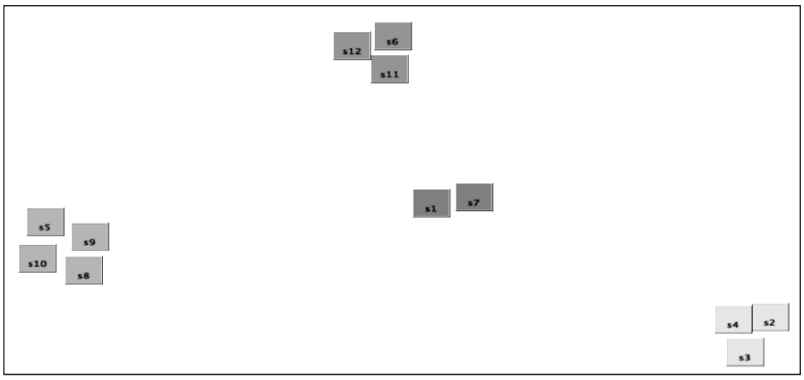
2.3.1 Task

Participants completed six cluster trials: 3 trials with *Alpha* speakers and 3 trials with *Betha* speakers (random order; non-repeating). Each trial was composed of 20 speech recordings. Participants were tasked with grouping the speech recordings into five cluster groups, where each cluster represented a unique speaker. Participants were required to classify all recordings into one of the five groups.

2.3.2 Materials

Throughout the study, participants wore AKG K702 headphones and navigated the TCL-LABX interface (Gaillard, 2009) on a desktop computer in the CEP-LPL computer laboratory. The intuitive interface allows users to select and move recordings in a two-dimensional space. Each recording is represented as numbered square, where a single click launches audio playback. To assign a recording to a group, users right-click on the square, which, in turn, creates a drop-down menu with different colour options. Fig. 1 illustrates a screenshot of the TCL-LABX interface.

Figure 1 – *Illustration of the TCL-LABX interface. Each coloured square represents a speech recording*



2.3.3 Data processing

Oftentimes used in machine learning, the Mathews Correlation Coefficient (MCC) is a measure of the quality of binary classifications. The MCC (Equation 1) was selected to determine how accurate the participants were at discriminating speakers, where *TP*, *TN*, *FP*, *FN* represent the selections that were "true positive", "true negative", "false positive" and "false negative", respectively. The mode speaker in each cluster was used to calculate the MCC mean and standard deviation for each speaker was taken.

(1) $$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$$

As an alternative method of analysing performance, we proposed the cluster *purity* metric, which identifies a different speaker per cluster in a trial (Equation 2). Unlike MCC, purity focuses only on maximising the total number of "true positive" responses per cluster. Purity values range between 0 and 1 (perfect clustering). We define purity as:

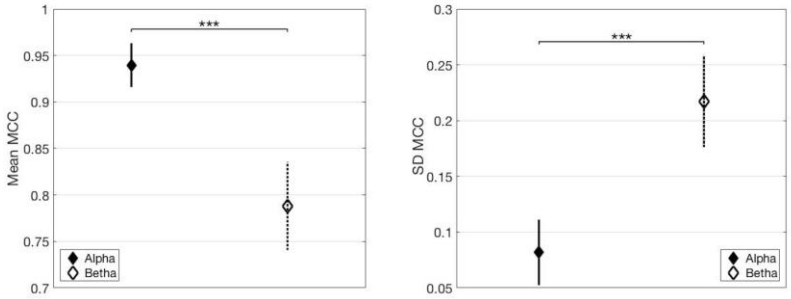
(2) $$purity(M) = \max_k \frac{1}{N} \sum_{m \in M} m \cap d_m^k$$

where M is a trial, m is a cluster in trial M, k is the number of speakers, d^k is the different combinations of unique speakers assigned to each cluster in trial M, and N is the number of speech recordings in trial M.
In addition to these metrics, for each speaker we calculated the CDS mean and standard deviation between her and the other speakers in her group.
To examine participant performance discriminating speakers, two-level nested ANOVA procedures were applied to MCC and purity mean and standard deviation metrics for groups with different speakers ($\alpha = 0.05$). Where main effects were detected, post-hoc Bonferroni-adjusted t-tests were carried out.

3. *Results.*

We found a main effect on groups for MCC mean $F_{1,240} = 32.12$, $p < 0.001$, $\eta_p^2 = 0.12$, and no significance between speakers within each group, $p > 0.05$. Post-hoc tests revealed the *Alpha* group had a higher MCC mean (0.94 ± 0.02) when compared to the *Betha* group (0.79 ± 0.05), $p < 0.001$ (Fig. 2-Left). Similarly we found a main effect on MCC standard deviation $F_{1,240} = 28.24$, $p < 0.001$, $\eta_p^2 = 0.11$, but again no significance between speakers within each group, $p > 0.05$. Post-hoc tests revealed the *Alpha* group had a lower MCC standard deviation (0.08 ± 0.03) when compared to the *Betha* group (0.2 ± 0.04), $p < 0.001$ (Fig. 2 -Right).

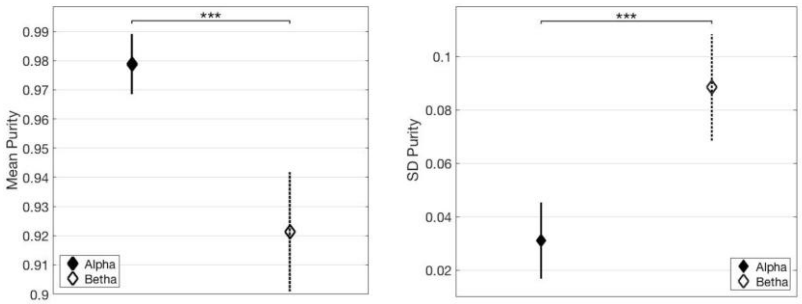
Figure 2 – *Group MCC means (Left) and standard deviations (Right). Diamonds and vertical lines represent means and standard error, respectively.* {***} *represents p < 0.001 with* $\alpha = 0.05$



In general we observed similar effects on purity metrics. We found a main effect on groups for purity mean $F_{1,240} = 24.63$, $p < 0.001$, $\eta_p^2 = 0.09$, and no significance between speakers within each group, $p > 0.05$. Post-hoc tests revealed the *Alpha* group had a higher purity mean (0.99 ± 0.01) when compared to the *Betha* group (0.92 ± 0.02), $p < 0.001$ (Fig. 3-Left.) Similarly we found a main effect on purity standard deviation $F_{1,240} = 21.32$, $p < 0.001$, $\eta_p^2 = 0.08$, but again no significance between speakers within each group, $p > 0.05$.

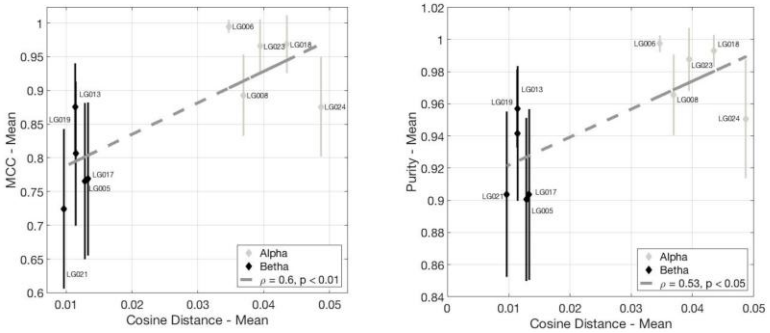
Post-hoc tests revealed the *Alpha* group had a lower purity standard deviation (0.03 ± 0.01) when compared to the *Betha* group (0.09 ± 0.02), $p < 0.001$ (Fig. 3 -Right).

Figure 3 – *Group Purity means (Left) and standard deviations (Right). Diamonds and vertical lines represent means and standard error, respectively. {***} represents $p < 0.001$ with $\alpha = 0.05$*



Next we examined whether our method of selecting and grouping speakers played a role in participant performance. Using the CDSs that were generated by our ASV system to make speaker group selections, we calculated the mean and standard deviation of difference between each speaker speech recording and the speech recordings of other speakers in its group. We then calculated the Pearson correlation coefficient to examine the relationships between participant performance and speaker CDS. The speaker CDS mean difference correlated to MCC mean at $\rho = 0.6$, $p < 0.01$ (Fig. 4-Left), whereas the speaker CDS standard deviation correlated to MCC standard deviation at $\rho = 0.52$, $p < 0.05$. We observed similar correlations between CDS and purity metrics, where means correlated at $\rho = 0.53$, $p < 0.05$ (Fig. 4-Right), and standard deviations at $\rho = 0.51$, $p < 0.05$.

Figure 4 – *Pearson correlation procedures applied to: speaker CDS and MCC means with $\rho = 0.6$, $p < 0.01$ (Left) and speaker CDS and Purity means with $\rho = 0.53$, $p < 0.05$ (Right). Diamonds and vertical lines represent means and standard error, respectively. Text refers to speaker identification*



4. *Discussion.*

Our findings revealed that participants were able to use the clustering interface to make discriminations based on their perceived differences between speech recordings. Participants performed at a relatively high level, as indicated by the mean and standard deviation of MCC and purity values, which suggests they found the interface easy to navigate and efficient to use. Our reported significant differences between groups suggest the importance of developing methods for selecting and grouping speaker stimuli.

We observed very little differences between the MCC and purity metrics. By using these two different metrics, our goal was to observe any differences between their assessments of participant performance and whether they correlated differently to our objective metrics (CDS). As we previously stated the traditional MCC metric evaluates performance based on binary responses, whereas the purity metric was more robust, as it adapted to the speech recording compositions in each cluster. The MCC metric is typically used in conventional speaker identification tasks and, as evidenced by our findings, it proved to be suitable for evaluating perceptual performance in a more open and flexible task. The purity metric, however, was developed for such an environment and provided a more detailed profile of each listener and their listening capacities and limitations. Our initial findings suggest the purity metric has promise, as it captured participant performance in a manner similar to traditional speaker identification metrics.

Looking more closely at the groups and the speakers composed in them, several observations can be drawn. Participants were more accurate with the *Alpha* group, which was expected, as it was mixed-gender and unbalanced (3 female, 2 male speakers). Acoustic speech features such as accent or vowel sounds were probably not the primary indices used by participants, who more likely relied on differences between F0 and vocal timbre. Our findings revealed that participants - overwhelmingly - found it easy to identify speaker LG006, as she had the highest mean MCC (0.95 ± 0.01) and purity (0.99 ± 0.01) metrics with very little deviations. Interestingly, we observed similar averages and variations between the remaining male and female speakers, where participants were more accurate with the latter gender. It is possible that, because there were no female speakers in the *Betha* group, participants were more sensitive to differences between them, which, in turn, improved their speaker identification accuracy.

Conversely, accuracy decreased when participants were presented *Betha* group speakers. We observed that, when compared to the *Alpha* speakers, there were smaller differences between the CDS means and standard deviations of *Betha* speakers. This of course was the motivating factor for selecting them from the PTSVOX corpus. Our results suggest that participants equally found the speakers to be quite similar, as indicated by the lower MCC means and higher standard deviations. All the *Betha* speakers were male, which suggests participants were required to be more sensitive in their listening in order to distinguish speakers. However, it was difficult to assess any inter- or intra-

speaker discrimination strategies employed by participants. A study by Baumann and Belin (2008) reported that naïve listeners used F0 and mean difference between F4 and F5 different to identify male speakers. Future research could be conducted to assess whether there were any significant differences between these acoustic features across speakers. As provided in Tab. 1, we observed that both LG005 and LG017 speakers come from the Rhône region, so it is possible that participants perceived a likeness between them, which might have led to confusion that affected accuracy. This point is of course difficult to claim, given the limited stimuli and tests. However, it does bring into discussion the subject of accents in speaker identification task, which has become a recent research focus in the domain of computational linguistics (Hannani, Russell, and Carey, 2013).

Turning to the results of our correlation procedures, we observed that an increase in speaker CDS mean lead to less accurate identifications, and, conversely, a decrease in speaker CDS standard deviation lead to an increase in speaker identification variability. These results suggest that ASV systems can be useful for preliminary speaker identification estimations by naïve listeners. A study by Gerlach et al. (2020) reported positive relationships between listener judgements and scores produced by ASV systems for both English and German language speakers. However, these findings contrast those reported in studies by Lindh and Eriksson. (2010) and Zetterholm, Elenius, and Blomberg (2004). Similarly, an important study by Park et al. (2018) found that not only did naïve listeners outperform an ASV system when completing a text-independent speaker discrimination task, but a weak correlation between human and machine performance. Taken these findings together suggest there are remain important differences in how humans and machines represent speakers.

The clustering task was designed to be more open in comparison to traditional perceptual speaker identification tasks. Both voice parades and pairwise comparisons restrict the modes in which users can express judgements. This differs from perceptual clustering tasks, which offer users dynamic engagements with speech materials. Rather than tasking listeners to judge whether a target speaker is present in a set (voice parade), the perceptual clustering method neutralises the concept of a "target" by asking them to organise a set of speech materials in terms of their likeness. It also provides naive listeners with a larger set of speech stimuli in which they could freely familiarise and group, which contrasts pairwise comparison method. While this study designed trials to include 20 speech stimuli, we have developed other studies with perceptual clustering tasks that include 12 (O'Brien, Meunier, and Ghio, 2021a) and 15 speech stimuli (O'Brien, Chanclu, Tomashenko, and Bonastre, 2021b). Moreover, a study by Lavan et al. (2021) found that a voice-sorting task provided familiar listeners with an advantage over unfamiliar listeners, which suggests listeners find any accessible information useful to evaluate speech materials.

Although these findings revealed promising results, there are still many factors to consider for future work, such as the number of stimuli per session and the number of different speakers per group. Five different speakers were selected as it reflected a number typically used in a visual lineup. However, in an effort to remain balanced, it was possible that participants were able to correctly deduce that there were four speech recordings per group. To combat any (potential) deductions, we might consider developing an unbalanced design, where sessions are composed of an unequal number of speaker speech recordings. By contrast, we might change the task's instructions, such that participants are asked to organise speech recordings into a maximum of four (or more) groups.

Unfortunately we were limited with the features we could measure with the interface. One feature that might offer additional insight is the number of times a participant listened to a speech recording, which could then be used to measure differences between participants and speakers. In addition, the number of times a speech recording was moved or classified might also support our analysis and provide us with a better parameter to measure the effects of the interface. These are some of the possibilities afforded by a clustering interface, which, when joined with other performance variables, could be used in combination for joint factorial analysis to better understand the interconnections between speaker discrimination performance and acoustic features that characterise speech.

5. *Conclusions.*

Our perceptual clustering method highlighted how naïve listeners performed at a high level, which correlated to scores produced by our ASV system. Because of its design the perceptual clustering method produced idiosyncratic responses for each listener. In comparison to more restrictive perceptual speaker identification tasks, the clustering method could be used to not only profile listeners and understand better their perceptive capacities, but also to eliminate speech recording outliers, e.g. if certain recordings were too easily (mis)identified by listener populations. The selection of foil speech materials could be done in this manner, where ASV systems first select materials and, subsequently, listeners evaluate speech recordings via a perceptual clustering task to identify any biasing speech characteristics.

Overall, our results demonstrate that naïve listeners were effective at using the perceptual clustering task to identify speakers. This research led us to develop two different parallel speaker discrimination studies. As part of the same *VoxCrim* project, the first study examined the effects of different perceptual speaker identification tasks with similar stimuli (O'Brien, Meunier, and Ghio, 2021a). This study centred on the voice parade, pairwise comparison, and clustering perceptual speaker identification tasks and examined the relationships between accuracy, task-dependent temporal features, and scores generated by an ASV system across and between tasks. Our findings from this study suggest

that each perceptual task has the capacity to deliver important information that naive listeners can use when identifying and discriminating speakers.
A second study that employed a perceptual clustering method was part of the "VoicePrivacy Challenge," collaborative, multi-objective project that aims to address questions surrounding speaker anonymization (Tomashenko, Srivastava, Wang, Vincent, Nautsch, Yamagishi, Evans, Patino, Bonastre, Noe, and Todisco, 2020). As part of this collaborative project, a study by O'Brien et al. (2021b) developed a more sophisticated clustering interface, which added some of the features mentioned above. Our goal was to examine the effectiveness of two different speaker anonymization systems. Like the current study, our findings showed participants were able to use the perceptual clustering method to link natural and anonymised speech recordings. Coupling these studies together, it is clear that a perceptual clustering method is robust and offers an innovative approach to studying voice perception.

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.

Bibliography

Baumann, O., Belin, P. (2008). "Perceptual scaling of voice identity: Common dimensions for different vowels and speakers," *Psychological research* 74: 110–20. doi:75310.1007/s00426-008-0185-z.

Blatchford, H., Foulkes, P. (2006). "Identification of voices in shouting," *International Journal of Speech, Language and the Law* 12(2): 241-254

Hollien, H., Foulkes, P. (2006). "Identification of voices in shouting," *International Journal of Speech, Language and the Law* 12(2): 241-254.

Belin, P., Bestelmeyer, P., Latinus, M., Watson, R. (2011) "Understanding Voice Perception," *British journal of psychology* 102: 711-25. doi:10.1111/j.2044-8295.2011.02041.x.

Boë, L., Bonastre, J-F. (2012) "L'identification du locuteur: 20 ans de témoignage dans les cours de Justice. Le cas du LIPSADON «laboratoire indépendant de police scientifique»", *Actes de JEP* 2012, Grenoble: 417-424.

Bonastre, J-F., Wils, F., Meignier, S. (2005) "Alize, a free toolkit for speaker recognition," *Acoustics, Speech, and Signal Processing* 1988. doi: 10.1109/ICASSP.2005.1415219.

Broeders, T., Amelsvoort, A. (1999) "Lineup Construction for Forensic Earwitness Identification: a Practical Approach," in *Proc. XIVth International Congress of Phonetic Sciences*: 1373-1376.

Chanclu, A., Georgeton, L., Fredouille, C., Bonastre, J-F. (2020) "PTSVOX: une base de données pour la comparaison de voix dans le cadre judiciaire," *Actes de 6e conférence conjointe Journées d'Études sur la Parole*, Nancy, FR: 73-81.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P. (2011) "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing* 19(4): 788–798.

Fleming, D., Giordano, B. L., Caldara, R., Belin, P. (2014). "A language familiarity effect for speaker discrimination without comprehension," in *Proc. National Academy of Sciences* 111(38).

Fredouille, C., Charlet, D. (2014) "Analysis of I-Vector framework for Speaker Identification in TV-shows," in *Proc. Interspeech 2014*, Singapour, Singapore.

Gaillard, P. (2009). Laissez-nous trier! TCL-LabX et les tâches de catégorisation libre de sons.

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 Communications*. doi:77810.1016/j.specom.2020.08.003.

Gravier (2021) Speech Signal Processing (SPro) Toolkit. https://www.irisa.fr/metiss/guig/spro.

Hannani, A., Russell, M., Carey, M. (2013) "Human and computer recognition of regional accents and ethnic groups from British English speech", *Computer Speech & Language,* 27(1): 59-74. doi:10.1016/j.csl.2012.01.003.

Harris, M., Gries, S., Miglio, V. (2014). "Prosody and its application to forensic linguistics," *Linguistic Evidence in Security, Law and Intelligence* 2. doi:10.5195/lesli.2014.12.

Hollien, H., Bahr, R., Künzel, H., Hollien, P. (2013). "Criteria for earwitness lineups," *Int. Jnl of Speech Language and the Law* 2: 143-153.

Hollien, H., Majewski, W., Doherty, E.T. (1982) "Perceptual identification of voices under normal, stress and disguise speaking conditions," *J. Phonetics* 10: 139-148.

Jenson, D., Saltuklaroglu, T. (2021). "Sensorimotor contributions to working memory differ between the discrimination of same and different syllable pairs," *Neuropsychologia* 159. doi:https://doi.org/10.1016/j.neuropsychologia.2021.788107947.

Jong, G., Nolan, F., McDougall, K., Hudson, T. (2015). "Voice lineups: A practical guide," in *Proc. ICPhS*.

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M. (2011). "i-vector Based Speaker Recognition on Short Utterances," in

Proc. Annual Conference of the International Speech Communication Association, *Interspeech 2001*. doi:10.21437/Interspeech.2011-58.

Kerstholt, J., Jansen, E., Amelsvoort, A., Broeders, T. (2006). "Earwitnesses: Effects of accent, retention and telephone," *Applied Cognitive Psychology* 20: 187-197. doi:10.1002/acp.1175.

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

Kreiman, J., Van Lancker Sidtis, D. (2011). Foundations of Voice Studies: An Inter-disciplinary Approach to Voice Production and Perception.

Larcher, A., Bonastre, J-F., Fauve, B., Lee, K, Levy, C., Li, H., Mason, J., Parfait, J-V. (2013) "ALIZE 3.0 - Open source toolkit for state-of-the-art speaker recognition," in *Proc. Of Interspeech* 2013, Lyon.

LaRiviere, C. (1971). "Some acoustic and perceptual correlates of speaker identification," in *Proc. 7th Int. Congress Phonetic Sciences*: 558-564.

Lavan, N., Burston, L., Garrido, L. (2018). "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices," *British Journal of Psychology* 110. doi:10.1111/bjop.12348.

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. doi:10.1016/j.cognition.2021.104780

Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T. (2016) "Speaker identification and clustering using convolutional neural networks," in *Proc. IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP):* 1-6, doi: 10.1109/MLSP.2016.7738816.

Lindh, J., Eriksson, A. (2010). "Voice similarity - a comparison between judgements by human listeners and automatic voice comparison, " in *Proc. FONETIK 2010*: 63-69.

Mattys, S. Davis, M., Bradlow, A., Scott, S. (2012). "Speech recognition in adverse conditions: A review," *Language and Cognitive* 27: 953-978.

Mühl, C., Sheil, O., Jarutyte, L., Bestelmeyer, P. (2017a). "The bangor voice matching test: A standardized test for the assessment of voice perception ability," *Behavior Research Methods* 50: 1-9. doi:10.3758/s13428-017-0985-4.

Mullennix, J., Ross, A., Smith, C., Kuykendall, K., Conard, J., Barb, S. (2011) "Typicality effects on memory for voice: Implications for earwitness testimony," *Applied Cognitive Psychology* 25: 29-34. doi:10.1002/acp.1635.

O'Brien, B., Meunier, C., Ghio, A. (2021a). "Presentation matters: Evaluating speaker identification tasks," in *Proc. Interspeech 2021*. doi:10.21437/Interspeech.2021-1211.

O'Brien, B., Tomashenko, N., Chanclu, A., Bonastre, J.-F. (2021b). "Anonymous speaker clusters: Making distinctions between anonymised speech recordings with clustering interface," in *Proc. Interspeech 2021*. doi:10.21437/Interspeech.2021-1588.

Öhman, L., Eriksson, A., Granhag, P. (2010). "Mobile phone quality VS. Direct quality: How the presentation format affects earwitness identification accuracy," *European Journal of Psychology Applied to Legal Context* 2.

Olsson, J. (2003). What is Forensic Linguistics? England: Nebraska Wesleyan University.

Park, S. J., 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," *The Journal of the Acoustical Society of America1* 44: 375-386. doi:10.1121/1.5045323.

Roebuck, R., Wilding, J. (1993). "Effects of vowel variety and sample length on identification of a speaker in a line-up," *Applied Cognitive Psychology* 7: 475–481.

Salako, S.E. (2006). "The Declaration of Helsinki 2000: Ethical Principles and the Dignity of Difference," *Medicine and Law* 2: 341-354.

Sloos, M., García, A-A., Andersson, A., Neijmeijer, M. (2019) "Accent-induced bias in linguistic transcriptions," *Language Sciences* 76 (101176). doi: 10.1016/j.langsci.2018.06.002.

Smith, H., Baguley, T., Robson, J., Dunn, A., Stacey, P. (2018). "Forensic voice discrimination: The effect of speech type and background noise on performance," *Applied Cognitive Psychology* 33. doi:10.1002/acp.3478

Smith, H., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., Stacey, P. (2020). "Voice parade procedures: optimising witness performance," *Memory* 28(1): 9312-17. doi:10.1080/09658211.9322019.1673427.

Stevenage, S. (2018). "Drawing a distinction between familiar and unfamiliar voice processing: Are view of neuropsychological, clinical and empirical findings, *Neuropsychologia* 116, 162–178. doi: 10.1016/j.neuropsychologia.2017.07.005. Epub 2017 Jul 8. PMID: 28694095

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," *Quarterly Journal of Experimental Psychology* 73. doi:10.1177/9391747021819888064.

Tomashenko, N., Srivastava, B., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J-F., Noe, P-G., Todisco, M. (2020). "Introducing the VoicePrivacy Initiative," in *Proc. of Interspeech 2020*: 1693-1697. doi:10.21437/Interspeech.2020-1333.

Zetterholm, E., Elenius, D., Blomberg, M. (2004). "A comparison between human perception and a speaker verification system score of a

voice imitation," in *Proc. 10th Australian International Conference on Speech Sciences & Technology*, Sydney.

BENJAMIN O'BRIEN - Aix-Marseille Université, CNRS, LPL, UMR 7309, France
benjamin.o-brien@univ-amu.edu

CHRISTINE MEUNIER - Aix-Marseille Université, CNRS, LPL, UMR 7309, France
christine.meunier@univ-amu.fr

ALAIN GHIO - Aix-Marseille Université, CNRS, LPL, UMR 7309, France
alain.ghio@univ-amu.fr

CORINNE FREDOUILLE - Laboratoire Informatique d'Avignon, Université d'Avignon, Avignon, France
corinne.fredouille@univ-avignon.fr

JEAN-FRANÇOIS BONASTRE - Laboratoire Informatique d'Avignon, Université d'Avignon, Avignon, France
jean-francois.bonastre@univ-avignon.fr

CAROLANNE GUARINO - Aix-Marseille Université, CNRS, LPL, UMR 7309, France
carolanne.guarino@gmail.com