Discriminating speakers using perceptual clustering interface

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

Discriminating speakers using perceptual clustering interface

Benjamin O'Brien[1], Alain Ghio[1],
Corinne Fredouille[2],
Jean-François Bonastre[2], Christine Meunier[1].

1 Aix-Marseille Université CNRS LPL, Aix-en-Provence, FR
2 Laboratoire Informatique d'Avignon, Avignon Université, FR
Corresponding author: benjamin.O-BRIEN@univ-amu.fr

INTRODUCTION The challenges facing listeners tasked to identify speakers are well documented.[1 2 3] In addition to providing listeners with high-quality speech recordings that accurately represent the speakers, the method of presentation itself is equally important.[4 5] Numerous perception studies have employed a binary approach, where participants are asked to judge whether two speech recordings are similar or different, as a way of examining the effects of such things as noise,[6] language familiarity,[7 8] and stimuli selection methods.[9] Oftentimes this requires numerous tests, which can be time-consuming for participants and experimenters. Moreover, there persists concern for memory bias, as a "fresh" voice is not equivalent to a voice that was presented in a previous binary test.

As an alternative, we proposed the development of perpetual *clustering* method, which is often employed in the domain of machine learning.[10 11] We theorized that this approach would allow users to better personalize their engagements with speech materials and organize their proximities in relation to their perceived likeness. In addition, it was more economical in terms of the number of trials required to assess a listener's ability to identify speakers.

In order to study the speaker discrimination performance of human participants using a perceptual clustering interface, it was important to organize and select stimuli based on how listeners perceive them as similar or different. Studies suggest listeners rely on a common set of acoustic features to identify speakers.[12 13] It is common in the development of automatic voice recognition and speaker identification system to extract MFCCs or LFCCs from speech recordings to train models. A popular trend in the field involves the transformation of these features into i-vectors, which are then used for training and testing, and has been shown to be quite accurate in identifying speakers.[14 15] Recent work has shown that Cosine Distance Scoring (CDS) with Within-Class covariance normalization (WCCM) is similarly effective while reducing the complexity of the task.[16] Our second objective was to examine the relationships between participant performance and the CDS scores generated from i-vectors.

METHODS Speech recordings were selected from the PTSVox database,[17] which included 24 francophone speakers (12 female and 12 male) who 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 (mean duration 118.96 ± 17.54 s). Using the ALIZE system,[18] 19 MFCCs, deltas, and delta-deltas (57 total features) were extracted and normalized from each recording and used to calculate i-vectors (200 dimensions). CDS were then calculated between each i-vector, whereupon the WCCM was computed over the entire set. Two groups of five speakers were selected, such that the *Alpha* group was composed of speakers with the greatest distance between them, whereas as the *Betha* group was composed of speakers with the smallest distance between them. For each speaker, twelve utterances were selected (120 recordings; mean duration 1.47 ± 0.51 s). Groups divided the six sessions, such that each session was balanced and composed of four different (non-repeating) chunks per speaker.

Twenty-four people, who self-reported good hearing, participated in our study. Their task was to group 20 speech recordings into five cluster groups, where each cluster represented a unique speaker. To do this, they used the TCL-LABX interface,[19] which allowed them to move recordings in a 2-D space and assign them to different clusters. They completed six sessions.

The Mathews Correlation Coefficient (MCC) was selected to determine how accurate the participants were at discriminating speakers (1), 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 and the MCC mean and standard deviation for each speaker was taken.

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

RESULTS To examine participant performance discriminating speakers, Two-level nested ANOVA procedures were applied to MCC mean and standard deviation for groups with different speakers. We found a main effect on groups for MCC mean $F_{1,240} = 32.92$, $p < 0.001$, $\eta_p^2 = 0.12$, 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 ± 0.20) when compared to the Betha group (0.8 ± 0.02), $p < 0.001$. Similarly we found a main effect on MCC standard deviation $F_{1,240} = 26.04$, $p < 0.001$, $\eta_p^2 = 0.1$, 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 ± 0.02) when compared to the Betha group (0.2 ± 0.02), $p < 0.001$ (Figure 1 – Top).

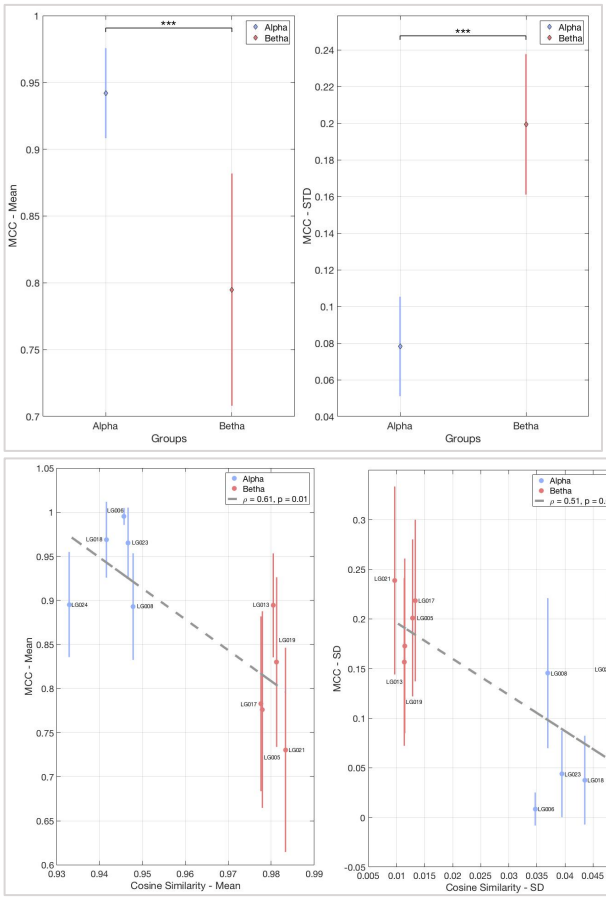

Figure 1-Top. Mean (Left) and standard deviation (Right) of participant Mathews Correlation coefficients (MCC) per group. Diamonds and vertical lines represent the means and standard errors, respectively. {***} signifies $p < 0.001$ with $\alpha = 0.05$. **Figure 1-Bottom**. Linear regression models comparing CDS and MCC metrics: mean (Left) $R^2 = 0.61$, $p = 0.01$, and standard deviation (Right) $R^2 = 0.51$, $p = 0.02$. Circles and vertical lines represent the means and standard errors, respectively. The text indicates speaker id.

Next we examined whether our method of selecting and grouping speakers played a role in participant performance. Using the CDS that were generated to make speaker group selections, we calculated the mean and standard deviation of difference between each speaker and the other speakers in its group. We then used these metrics with linear regression models to examine whether they could be used to estimate participant performance discriminating speakers. The speaker CDS mean difference estimated the MCC mean at $R^2 = 0.61$, $p = 0.01$, whereas the speaker CDS standard deviation estimated the MCC standard deviation at $R^2 = 0.51$, $p = 0.02$ (Figure 1-Bottom).

DISCUSSION This study demonstrated that users were able to use a 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 MCC values, which suggests they found the interface easy to navigate and efficient to use. In addition, the significant differences between groups also underscore the importance of developing methods for selecting and grouping speakers. We observed that as the CDS mean increased, participants were less accurate discriminating speakers, and, conversely, as the CDS standard deviation decreased, participants showed greater variability. These findings have led us to develop a new study to compare the effects of presentation on users performing speaker discrimination tasks with similar speaker stimuli.

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

References

[1] Cambier-Langeveld, T., Rossum, M., Vermeulen, J. (2014). Whose voice is that? Challenges in forensic phonetics.
[2] Mattys, S. Davis, M., Bradlow, A., Scott, S. (2012). Speech recognition in adverse conditions: A review. Language and Cognitive Processes - LANG COGNITIVE PROCESS. 27. 953-978.
[3] Nolan, F. (2001). "Speaker identification evidence: its forms, limitations, and roles," Law and Language: Prospect and Retrospect.
[4] 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 », 417-424. Actes de la conférence conjointe JEP-TALN-RECITAL 1: JEP, Grenoble.
[5] Hollien, H., Bahr, R., Künzel, H., Hollien, P. (2013). "Criteria for earwitness lineups," International Journal of Speech Language and the Law 2, 143–153.
[6] 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.
[7] Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language familiarity effect for speaker discrimination without comprehension. Proc National Academy of Sciences, 111(38), 13795–13798
[8] Levi, S. V., & Schwartz, R. G. (2013). The development of language specific and language-independent talker processing. Journal of Speech, Language, and Hearing Research, 56(3), 913–920.
[9] Mühl, C., Sheil, O., Jarutytė, L. et al. (2018) The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. Behav Res 50, 2184–2192.
[10] Kinnunen, T. and Kilpeläinen, T. (2000). Comparison of clustering algorithms in speaker identification. Proc. IASTED Int. Conf. Signal Processing and Communications (SPC). 222-227.
[11] Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. 1-6.
[12] LaRiviere, C. (1971). "Some acoustic and perceptual correlates of speaker identification," Proc 7th International Congress of Phonetic Sciences: 558–564.
[13] Roebuck, R., and 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.
[14] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P. (2011) Front-End Factor Analysis for Speaker Verification, in IEEE Transactions on Audio, Speech, and Language Processing, 19(4), pages 788–798.
[15] Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M. (2011). i-vector Based Speaker Recognition on Short Utterances. Proc International Speech Communication Association, INTERSPEECH.
[16] Fredouille, C., Charlet, D. (2014) Analysis of I-Vector framework for Speaker Identification in TV-shows. Interspeech, Singapour, Singapore.
[17] Chanclu, A., Georgeton, L., Fredouille, C., Bonastre, J-F. (2020) PTSVOX:

une base de données pour la comparaison de voix dans le cadre judiciaire. 6e conférence conjointe Journées d'Études sur la Parole, Nancy, FR. pp.73-81.
[18] Larcher, A., Bonastre, J-F., Fauve, B, et al. (2013) "ALIZE 3.0 - Open source toolkit for state-of-the-art speaker recognition," Proc. of Interspeech, Lyon.
[19] Gaillard, P. (2009). Laissez-nous trier ! TCL-LabX et les tâches de catégorisation libre de sons.