VOICE DISTINCTIVENESS: AN INVESTIGATION OF THE ROLE OF SPEAKERS' POSITION IN A POPULATION WITH RESPECT TO f0

Kirsty McDougall, Alice Paver and Francis Nolan

University of Cambridge
kem37|aep58|fjn1@cam.ac.uk

ABSTRACT

The role played by f0 in listeners' assessment of voice distinctiveness is investigated. In Experiment 1, listeners judged the (dis)similarity of low-, medium-, and high-pitched voices selected from an accent- and demographic-matched population. Listeners' judgements tended to cluster speakers together according to their position in a population distribution for f0, yet the tightness of these clusters varied, with members of the high-pitched group being consistently judged as most different from each other. This suggests that other phonetic dimensions are likely to be relevant as well as pitch. Experiment 2 collected similarity judgements of the same stimuli resynthesised with pitches shifted between the same low, medium and high positions in the population distribution. Results show that shifting the pitch of stimuli did not lead to a significant change in judgements, again indicating that more than pitch alone is driving the judgements phonetically. Implications for earwitness identification are discussed.

Keywords: voice distinctiveness, voice similarity, fundamental frequency, earwitness, voice parades

1. INTRODUCTION

In criminal cases where a witness has heard but not seen a perpetrator, a voice parade may be conducted to see whether the witness is able to pick out the voice of a suspect from a line-up of voice samples. Voice parade studies such as [4, 11] for familiar listeners and [7, 14] for unfamiliar listeners show certain target speakers being identified more readily than others: voices which sound more distinctive to a listener are presumably more likely to be recognised, yet the phonetic underpinnings of voice distinctiveness are not well understood.

[6] shows a correlation between listeners' judgements of voice distinctiveness and an acoustic metric combining f0, formants and harmonics-to-noise ratio. However, this study uses isolated utterances of *had* only, does not appear to account for demographic variation other than sex, and does not unpack the detail of the combined acoustic measure.

A study by Sørensen [13] focusses on the role of mean f0 in Danish listeners' ability to recognise unfamiliar male speakers within a group homogeneous for accent. Using a population distribution of mean f0, her experiment compared recognition of target speakers with centrally positioned mean f0 values with those with mean f0 values from the upper or lower tails of the distribution. Mid f0 voices were recognised correctly in 56% of trials, improving to 74% for high or low f0 voices. Sørensen interprets this as showing that mid f0 voices are harder to remember and recognise than those with a more extreme mean f0. This may be the case, but further evidence is needed; it is not clear how many different target voices were used and it is possible that further individual factors were at play. A crucial question arising from Sørensen's work is whether the relevant population for a voice parade is the population of foil voices used in the parade or the whole population.

The present study investigates whether, for a population homogeneous for accent/demographic characteristics, the distinctiveness of a voice is linked with its position in the population's distribution of mean f0 values. It tests whether speakers at the extreme ends of the f0 distribution are judged as sounding more distinctive from one another than speakers in the centre of the distribution through two experiments. In Experiment 1, listeners judge the similarity of pairs of speakers from within and between the lowest, highest and central parts of a population distribution for f0. In Experiment 2, listeners judge the similarity of the same group of speakers whose speech has been resynthesized across low, medium and high f0 positions.

2. EXPERIMENT 1

2.1. Method

2.1.1. Speakers

The *DyViS* database [9] provides a population of 100 matched-demographic speakers: male, aged 18-25 years, Standard Southern British English accent. Using 3-5 minutes net speech per speaker from a semi-spontaneous telephone call task (*DyViS* Task 2), the mean f0 was calculated for each speaker [5] using *Praat* [3]. f0 values were converted from Hertz to semitones (st, base 50Hz) to reflect the perceptual

scaling of pitch [8]. 12 speakers were selected, four from each of the lowest (speaker L1: mean f0 8.56st, L2: 8.77st, L3: 9.18st, L4: 9.38st), highest (H1: 16.7st, H2: 17.2st, H3: 17.3st, H4: 17.5st), and central (M1: 12.7st, M2: 13.00st, M3: 13.2st, M4: 13.5st) parts of the mean f0 distribution of the *DyViS* population. Each voice was reviewed auditorily to confirm that it did not sound particularly unusual (other than its pitch) before selecting it for the group.

2.1.2. Materials and experimental design

For each speaker, two three-second samples (U1 and U2) were created using short sections from the Task 2 recordings. Samples were controlled for content between speakers, taken from discussion of the same events prompted by the Task 2 scenario.

A similarity judgement task was built and hosted on Gorilla [1]. Each speaker was paired with himself and all other speakers, creating 78 pairings. A stimulus was prepared for each pairing, containing U1 and U2 separated by a silence of one second. Participants were randomly and evenly assigned to one of two groups, whereby in group 1 voice A appeared first, then voice B, and vice versa in group 2. Listeners rated the (dis)similarity of the two voices in each stimulus pairing on a Likert scale from 1 (very similar) to 9 (very dissimilar), having been asked to take into account voice quality and accent, and to ignore any meaningful content of the utterances, i.e. to focus on the sound of the voice. After six practice judgements, where listeners heard all 12 voices and became familiar with the task, they heard all pairings in a randomised order, with five evenly spaced rest breaks amongst the stimuli. The experiment took on average 17 minutes to complete.

2.1.3. Listener participants

35 participants aged 18-40 years (18 female, 16 male, 1 undisclosed) were recruited using the online recruitment platform Prolific.co. They lived and had spent most of their lives in the UK. They were native British English speakers with no reported hearing difficulties. Participants undertook a test [15] to ensure they were wearing headphones.

2.2. Results

Multi-dimensional scaling [12] was applied to the Likert scale judgements to reduce the many pairwise distances in a smaller number of dimensions. The relative positions of the 12 speakers according to the first two (perceptually most important) MDS dimensions are shown in Figure 1. Speakers in the same pitch group are clustered close together, particularly along dimension 1, with some overlap for

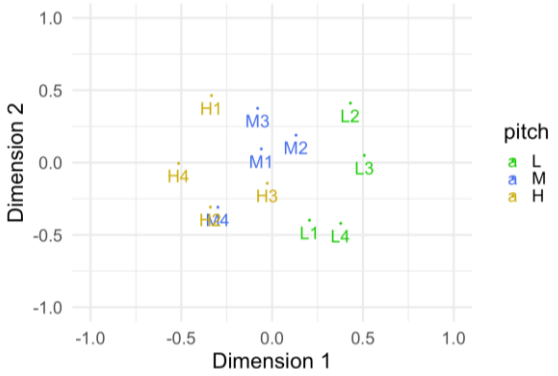

Figure 1: Scatterplot of the 12 speakers' locations on the first two dimensions (of five) produced by MDS using listeners' judgements of the speaker pairings [S-stress = 0.08299, Dispersion Accounted For (D.A.F.) = 0.97851]

the high- and medium-pitched speakers. Medium group speakers fall centrally in the display, with the high group overlapping and to the left of the overall collection of speakers, and the low group on the right side of the collection. In other words, along Dimension 1, the speakers are distributed approximately according to their pitch relationships, but neither the high nor low groups are positioned in an extreme location.

Regarding within-pitch-group comparisons, the individual members of the high group were judged to be more different from one another than the members within the low and medium groups, with H1 to H4 being spread further apart than L1 to L4 and M1 to M4. The boxplots in Figure 2 illustrate this; the median Likert scale score for the high group was 7, while the low and medium groups each had a median Likert scale score of 3. A Kruskal-Wallis test shows pitch group has a significant effect on Likert scores ($\chi^2(2) = 127.71$, $p < 0.0001$). A post-hoc Dunn test shows that the high group is significantly different from both the low group ($p < 0.0001$) and the medium group ($p < 0.0001$), whilst medium and low are also significantly different from one another ($p = 0.0186$).

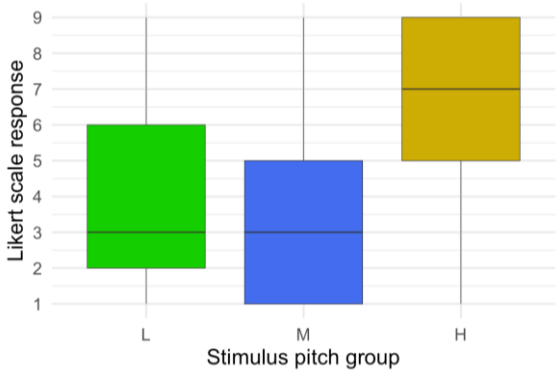

Figure 2: Boxplots of the Likert scale listener responses for within-group comparisons for each pitch group.

2.3. Interim summary

Voice (dis)similarity judgements show some clustering among speakers according to their pitch groups. However, the extent to which speakers are spread apart within each pitch group differs depending on the group. The clustering adds evidence that pitch plays a role in similarity judgements, while the spreading highlights the possibility that other speech dimensions are contributing. In particular, the speaker pairings within the high group are spread further apart in their Likert scale judgements than pairings in the other two pitch groups: could this be due to the high speakers differing more widely on speech dimensions other than pitch? To investigate this further, a second experiment was conducted in which similarity judgements were collected for the same speech stimuli which had been resynthesised after pitch-shifting. If speech dimensions other than pitch are driving the judgements, the high group speakers should always be the most different from one another even if they are resynthesised to have a low or medium pitch.

3. EXPERIMENT 2

3.1. Method

3.1.1. Materials and experimental design

The speech stimuli were all resynthesised so that there were three versions of each stimulus in low, medium and high pitch. For each stimulus, two of these resynthesized versions were a pitch-shifted version of the stimulus. The third was a resynthesised version of the stimulus at the same pitch as the original, undertaken to make sure that all stimuli sound equally 'unnatural' as a result of their resynthesis. The resynthesis process included all possible options as in Table 1. Stimuli under-went pitch-shifting in *Praat* using the 'Manipulate' and 'Shift pitch frequencies…' tool to shift the pitch contour by the appropriate number of semitones for the new median pitch value. (Note while speakers were chosen using *mean* f0 values, *Praat* uses *median* f0 for pitch-shifting). Speakers occupied the same relative position within a new pitch group, i.e., speaker L1 became speaker LM1 when shifted to medium pitch and LH1 when shifted to high pitch. Pitch values for the transformation were determined by the corresponding speaker in the new pitch group; for example, when being shifted to the new medium pitch group, speaker L1 was manipulated from his original median f0 of 8.56st to 12.7st, the original f0 of speaker M1. Stimuli were then evaluated to ensure naturalness after resynthesis. The same experimental set-up as for Experiment 1 was used, but in Experiment 2 listeners heard one of three different combinations of resynthesised stimuli. All listeners listened to the original 12 speakers whose stimuli had undergone resynthesis.

Resynthesis	Code
Low → Low	LL
Low → Medium	LM
Low → High	LH
Medium → Low	ML
Medium → Medium	MM
Medium → High	MH
High → Low	HL
High → Medium	HM
High → High	HH

Table 1: Pitch shift patterns used in Experiment 2.

The 12 speakers were heard in three pitch groups, Low, Medium and High, with four speakers in each. The combinations of resynthesised stimuli played to listeners are shown in Table 2. 20 listeners were

	Stimuli Group		
Stimuli Combination	Low-pitched	Medium-pitched	High-pitched
Combination 1	LL	HM	MH
Combination 2	HL	MM	LH
Combination 3	ML	LM	HH

Table 2: The three combinations of resynthesised stimuli, each heard by a separate group of listeners.

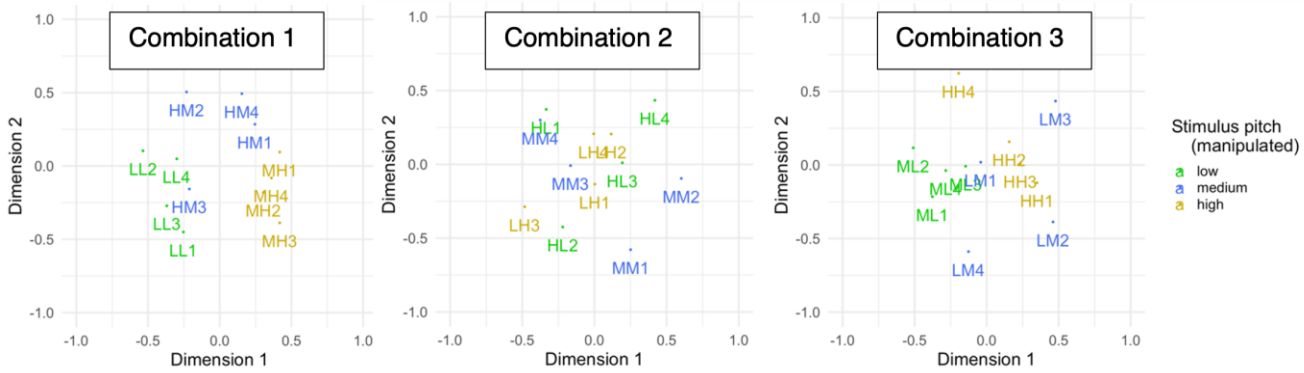

Figure 3: Scatterplots of the 12 speakers' locations on the first two dimensions (of five) yielded by MDS using judgements of the resynthesised stimuli for Combinations 1 (S-stress 0.08188, D.A.F. = 0.97865), 2 (S-stress .08789, D.A.F. = 0.97618) and 3 (S-stress .08477, D.A.F. = 0.97819).

randomly allocated to each stimulus combination.

3.1.2. Listener participants

60 participants (30 male, 29 female, 1 undisclosed), different individuals from Experiment 1 with the same characteristics, were recruited using Prolific.co.

3.2. Results

Scatterplots of the first two MDS dimensions for each stimuli combination are shown in Figure 3. In Combinations 1 and 3, speakers form relatively clear clusters according to their new pitch groups. In Combination 2, there is much more overlap among pitch groups, perhaps due to the fact that for the shifted speakers the pitch shift was the most extreme, i.e. low to high, and high to low.

Considering the relative positions of speakers, the shifted-to-medium group speakers (speakers labelled '…M') fall relatively centrally in each of the three graphs, but only for Combination 1 is there a sense of a low-medium-high continuum along Dimension 1, similar to that seen in Figure 1 for the natural stimuli. For Combination 3, the shifted-to-high pitch cluster largely falls within the datapoints of the shifted-to-medium cluster, while the shifted-to-low cluster is to the left of the overall collection of speakers. Meanwhile the display for Combination 2 does not show clear pitch-group-based clusters.

Figure 4 shows boxplots of the Likert scale judgements for the pitch groups in each stimuli combination. Reading each of the top, middle, and bottom panels of Figure 4 from left to right, it is clear that the originally H speakers (at right) still yielded the highest Likert scores regardless of f0 manipulation. It can be inferred that while pitch is clearly important, dimensions in addition to pitch are driving the similarity/difference judgements of these speakers. Reviewing the stimuli auditorily shows that

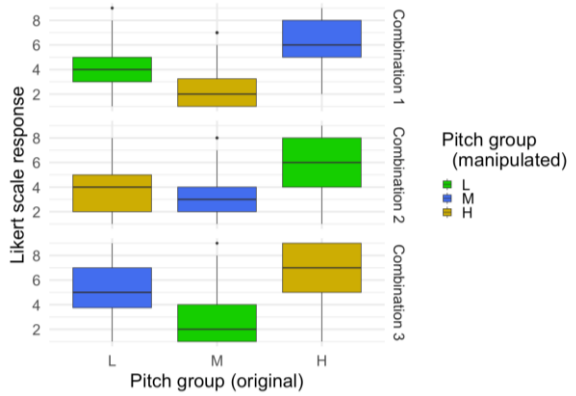

Figure 4: Box plot of the Likert scale listener responses for within-group comparisons for the low, medium and high pitch groups in Combinations 1 (top panel), 2 (middle panel) and 3 (bottom panel)

the originally high-pitched speakers have a greater range of voice qualities when compared with the originally low- and medium-pitched groups.

A Kruskal-Wallis test showed that pitch-shift status had no significant effect on Likert scores for all the within-pitch-group comparisons combined ($\chi^2(2) = 0.187$, $p = 0.665$). This test was repeated for the individual pitch groups, confirming that pitch-shift status had no significant effect on Likert scores for the originally low-pitched ($\chi^2(1) = 2.421$, $p = 0.1197$), medium-pitched ($\chi^2(1) = 3.741$, $p = 0.0531$), or high-pitched ($\chi^2(1) = 0.983$, $p = 0.3215$) groups.

4. DISCUSSION AND CONCLUSION

When listeners judged the (dis)similarity of all pairings among a group containing speakers of extreme low, extreme high, and mid mean f0 values, pitch clearly played a role, with speakers from the same pitch group clustering together, and an apparent 'low-mid-high pitch continuum' forming in correspondence with the most important MDS dimension [cf. 2, 10]. Neither the high or low groups were positioned remotely from other speakers on the MDS scatterplot, so Sørensen's contention that speakers with an extreme mean f0 value are perceived as more distinctive (and therefore more memorable) is not directly refuted, yet it is not unqualifiedly confirmed. The MDS scatterplots also exhibited varied extents of spreading of the speakers' locations within each pitch group, indicating that other phonetic dimensions are likely to be relevant.

Collecting (dis)similarity judgements for stimuli which had been resynthesised across the three pitch groups confirmed that while pitch is relevant, a more complex phonetic picture is at play. The within-pitch-group spread patterning observed for the original stimuli (larger spread of the H group speakers) was once again present and echoed in an auditory review, particularly with respect to voice quality.

The immediate implications of this study for voice parade construction are that a consideration of f0 is essential when selecting foils for comparison with a suspect, and that particular care should be taken in cases where the suspect's mean f0 is at the low or high extreme of the distribution for the relevant population, due to listeners' sensitivity to this feature. However, judgements of (dis)similarity are influenced by more than f0 alone [cf. 6], and further research is needed to uncover the complexity of the relationship between f0 and other phonetic features and speaker distinctiveness in order to better understand both what makes certain voices sound distinctive and the implications for earwitness procedures of the degree of perceived distinctiveness of a suspect's voice.

5. ACKNOWLEDGEMENTS

This research was supported by the UK Economic and Social Research Council as part of the project Improving Voice Identification Procedures (IVIP), reference ES/S015965/1. Additional funding was provided by the Isaac Newton Trust.

5. REFERENCES

[1] Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 31(1), 388-407.

[2] Baumann, O., Belin, P. 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research* 74, 110-120.

[3] Boersma, P., Weenink, D. 1992-2023. PRAAT: Doing phonetics by computer. [Computer program]. http://www.praat.org/.

[4] Foulkes, P., Barron, A. 2000. Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics* 7(2), 180-198.

[5] Hudson, T., de Jong, G., McDougall, K., Harrison, P., Nolan, F. 2007. F0 statistics for 100 young male speakers of Standard Southern British English. In: Trouvain, J., Barry, W.J. (eds), *16th International Congress of Phonetic Sciences, Saarbrücken, August 6-10, 2007, Proceedings.* Universitaet des Saarlandes, 1809-1812

[6] Latinus, M., McAleer, P., Bestelmeyer, P.E.G., Belin, P. 2013. Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology* 23(12), 1075-1080.

[7] McDougall, K., Nolan, F., Hudson, T. 2015. Telephone transmission and earwitnesses: performance on voice parades controlled for voice similarity. *Phonetica* 72, 257-272.

[8] Nolan, F. 2003. Intonational equivalence: an experimental evaluation of pitch scales. In: Solé, M. J., Recasens, D., Romero, J. (eds), *15th International Congress of Phonetic Sciences, Barcelona, August 3-9, 2003, Proceedings.* Causal, 771-774.

[8] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2009. The *DyViS* database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16(1), 31-57.

[9] Nolan, F., McDougall, K., Hudson, T. 2011. Some acoustic correlates of perceived (dis)similarity between same-accent voices. In: Lee, W.-S., Zee, E. (eds), *17th International Congress of Phonetic Sciences, Hong Kong, August 17-21, 2011, Proceedings.* City University of Hong Kong, 1506-1509.

[10] Rose, P., Duncan, S. 1995. Naive auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics*, 2(1), 1-17.

[11] Schiffman, S.S., Lance Reynolds, M., Young, F.W. 1981. *Introduction to Multidimensional Scaling: Theory, Methods, and Applications*. Academic Press.

[12] Sørensen, M.H. 2012. Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech, Language and the Law* 19(2), 145-158.

[13] Thompson, C.P. 1985. A language effect in voice identification. *Human Learning* 4, 19-27.

[14] Woods, K.J.P., Siegel, M., Traer, J., McDermott, J.H. 2017. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception and Psychophysics* 79(7), 2064-2072.