WHAT CAN VOICE LINE-UPS TELL US ABOUT VOICE SIMILARITY?

Renata R. Passetti[1], Sandra Madureira[2], Plínio A. Barbosa[3]
[1]Postgraduate Program in Linguistics, UFSCar, Brazil, [2] Graduate Program in Applied Linguistics and Language Studies, PUCSP, Brazil, [3]Dep. of Linguistics, UNICAMP, Brazil
re.passetti@gmail.com, madusali@pucsp.br, pabarbosa.unicampbr@gmail.com

ABSTRACT

This study considers the outcomes of a voice line-up experiment to shed light on the acoustic-prosodic features related to voice similarity. Data analysis comprises a similarity function based on acoustic parameters, and a multidimensional technique based on perceptual evaluation of voice quality. Target and foil stimuli are compared in terms of a set of 18 prosodic-acoustic parameters. The voice quality settings of the target, the most chosen foil and the non-chosen foil are also analysed according to the Voice Profile Analysis (VPA) and contrasted by a multidimensional scaling technique. The timing interval between stimuli presentations had great impact on voice identification correctness. The f0 parameters regarding melody were influential for establishing correct voice identifications. Voice quality settings were relevant regarding to false alarms, and voice dynamics settings for correct stimuli rejections. Results also highlight the importance of segmental and speaking tempo parameters to perceived voice similarity.

Keywords: voice line-up, voice similarity, VPA, similarity function

1. INTRODUCTION

Voice similarity is a pivotal topic in forensic phonetics. Knowing the phonetic features which can make voices sound similar is of special interest for forensic tasks concerning earwitness evidence. In such situations, voice similarity can be exploited both to improve procedures for preparing voice line-ups as well as to ensure fair identification when conducting them [1].

Regarding voice line-up construction, the analysis of voice similarity is crucial for selecting foil speakers, i.e., volunteers whose speech samples will be added to the voice line-ups. According to a set of guidelines on voice line-ups developed for application in the Dutch forensic context [2], the recording of foils should meet the description of the suspect's and perpetrator's voices in terms of physiological, prosodic, and sociolinguistic features.

Another key factor for voice similarity is related to the kind of parameters which may play a role on its perceptual assessment. This question was addressed by Lindh [3] in a study that compared the outcomes of a perceived voice similarity test to those of a line-up experiment. The results showed voice similarity judgments were partially explained by similarities in speaking tempo parameters, such as articulation rate and pausing measurements.

In a follow-up study, Lindh and Eriksson [4] compared the outcomes of voice similarity judgments made by listeners to those of automatic systems to explore convergence and divergence between them. The results pointed out to the influence of linguistic factors, such as speaking style, on listeners' judgments, which do not seem to play a role in the automatic systems analysis. Furthermore, the authors suggest that when performing line-ups tasks, listeners may also pay attention to other linguistic factors, such as pronunciation and voice quality.

The acoustic correlates of perceived voice similarity were also topic of a study by Nolan et al. [5]. To avoid interference of linguistic factors on listeners' perception, speech samples from the same dialect were used in a paired comparison test. Correlations between the perceptual dimensions of similarity judgments and the set of acoustic variables have shown the major importance of fundamental frequency in perceived voice similarity. Other acoustic parameters in order of importance from higher to lower are F3, F2 and F1.

The present study considers issues raised in the literature on voice similarity and uses the outcomes of a voice line-up experiment to shed light on the prosodic features related to perceived voice similarity. Some of our research questions are: What prosodic cues may play a role in listeners performance on voice line-ups? In misleading identifications, are there prosodic similarities between the target and the most chosen foil voices? Why were some foils not chosen?

The methodological approach comprises perceptual and acoustic analysis of speech samples, which were used in perceptual tasks performed by attendants in a workshop on voice similarity held by the authors.

2. METHODOLOGY

2.1. Voice database

The speech samples used in the perceptual experiment were selected from the Corpus Forense do Português Brasileiro (henceforth CFPB). This corpus

has been compiled by the Audiovisual and Electronics Section of the Criminalistics Institute of the Brazilian Federal Police and consists of text-reading and semi-spontaneous recordings of officials from Brazilian law enforcement agencies.

At the time of the study the corpus contained 280 recordings from male speakers and 70 from female speakers, aged between 19 and 64 and from five regions in Brazil (North, Northeast, Central-West, Southeast and South). All the CFPB recordings are recorded at a sampling rate of 44.1 kHz.

2.2. Voice line-up experiment

2.2.1. Line-ups

The perceptual experiment comprised four voice line-ups. The line-ups were constructed following the guidelines proposed by [2] and consisted of the target voice randomly arranged with five other distractor voices.

The selection criteria of the 24 voices (six for each line-up) were mainly based on the availability of the samples in the CFPB. Therefore, we selected semi-spontaneous recordings of male speakers originally from Brazilian states with more than 10 samples at the corpus.

The chosen states were Ceará (CE) in the Northeast, São Paulo (SP) and Rio de Janeiro (RJ) in the Southeast, and Paraná (PR) in the South. For each state, the target and distractor voices were selected based on their similarity regarding the mean of f0, F1 and F2. This analysis was performed automatically by a Praat script developed by the third author.

The line-up stimuli lasted twenty seconds, while the target stimuli lasted forty seconds. We ensured that the content of the target speakers' speech in their first presentation was different from the content of their speech within the line-up since similar content could influence listeners' choices.

Finally, another Praat script was developed to automatically arrange the stimuli along the line-ups and normalizing them to the same level of intensity. Moreover, this script added a five-second-long silent pause between the stimuli.

The position of the target stimulus in each line-up is shown in Table 1.

Line-up	Target position
Ceará (CE)	#1
São Paulo (SP)	#3
Rio de Janeiro (RJ)	#5
Paraná (PR)	#5

Table 1: Position of the target stimuli in the line-ups.

2.2.2. Experimental procedure

The workshop practices were held over two days separated by a one-week interval. Two voice line-ups were performed in the first day, and two in the second day. Therefore, the experimental procedure consisted of running two voice line-ups (from CE and SP) immediately after the target voice exposure, one voice line-up (from RJ) one week after the target voice exposure, and one voice line-up (from PR) 30 minutes after the target voice exposure. The target and line-up voices were played only once, and the attendees registered their choices on an online form.

2.2.3. Listeners

From the 21 participants who attended the workshop, the sixteen who attended it on both days were included in the present study.

They were 11 men and 5 women, aged between 20 and 64 years (mean of 34 years). All of them were higher educated (completed or ongoing) in speech-related courses, and three of them were forensic experts. None of them reported cognitive or hearing impairments.

2.3. Voice similarity analysis

2.3.1. Acoustic analysis

In order to verify whether prosodic-acoustic similarity reflects the listeners' choice, the third author implemented a function in R that computes this similarity according to three different criteria. Before that, the audio files were segmented into chunks separated by silent pauses.

Acoustic measures were computed with the Praat script "Prosody Descriptor Extractor" [6] on all chunks. This script extracts twelve fundamental frequency (f0) descriptors, two intensity descriptors and four voice quality descriptors. Among them, the following were significant in this study: f0 median, f0 semi-amplitude between quartiles, f0 minimum, f0 maximum, standard deviations of f0 local peak values, mean f0 peak rate, mean f0 peak bandwidth, mean f0 rates of rises and falls, spectral emphasis [7], Harmonic-to-Noise ratio (HNR) and jitter.

The similarity function gives the order of foils similar to the target starting from the most similar to the less similar. Three criteria of computing the distance between the target and a foil are used: (1) the first one computes the sum of absolute errors between target and a foil for each parameter where the error is computed as the difference between the medians of the corresponding parameters divided by the median of that parameter in the target voice; (2) the second one computes the sum of absolute errors between

target and a foil for each parameter where the error is computed as the difference between the medians of the corresponding parameters divided by the pooled standard-deviation of target and foil for that parameter; and (3) the third criterion arranges the foils in descending order according to the number of parameters with minimum error as established by criterion 1.

2.3.2. Perceptual analysis

In order to verify the role of voice quality (henceforth VQ) and voice dynamics (henceforth VD) settings on listeners' performance, the first and second authors assessed the target, the most chosen foil and the non-chosen foil stimuli for the SP, RJ and PR line-ups by means of the Vocal Profile Analysis (VPA) [8]. The CE line-up stimuli were discarded from this analysis since all listeners correctly chose the target's stimulus.

After having evaluated the stimuli separately, the first and second authors came up with an agreed version of the VPA for each of those stimuli. Then, the agreed versions of the VPA were subjected to the Multidimensional Scaling technique (MDS) [9, 10], which was performed by means of the R *smacof* package [11, 12]. This analysis was carried out in two steps: firstly, the VQ settings (i.e. vocal tract, muscular tension, and phonation features), and secondly, the VD settings (i.e. pitch, loudness, and speech rate features) as input parameters.

3. RESULTS AND DISCUSSION

3.1. Voice line-up experiment

The number of choices for each stimulus in the line-ups is shown in Figure 1. In all the line-ups, the target voices were the most correctly identified ones.

The better performance of the CE line-up can be explained by a combination of the latency time after the target voice exposure (immediately after it) and its position in the line-up (first position). The RJ line-up, on the other hand, had the worst performance, since only six listeners correctly chose it. This score can also be explained by the latency time of one week between the voice target exposure and the line-up running.

Some stimuli were not chosen by anyone. This was the case for all stimuli except the first (i.e. the target) in the CE line-up, the fifth stimulus in the SP line-up, the sixth in the RJ line-up and the second in the PR line-up.

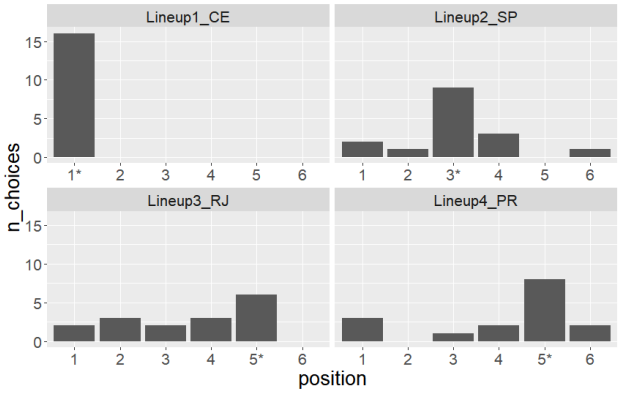

Figure 1: Number of choices for each stimulus position in the line-ups. Target positions are identified with an asterisk.

3.2. Acoustic analysis

The use of criterion 3 for the similarity function was the best predictor of the listeners' choices. For the purposes of analysis, in each line-up the two most similar foils to the target have been considered.

For the SP and RJ line-ups the function output predicted the target as the first choice. For the PR line-up, the target was the second function choice (foil 2, PR_2, being the first choice). The parameters with a minimum error according to criterion 3 which were common to the two correct choices were f0 semi-amplitude between quartiles, f0 maximum, mean f0 rates of rises and falls, and voice quality parameters (spectral emphasis for the SP line-up, and HNR and jitter for the RJ line-up). For the PR line-up, the parameters with a minimum error were f0 median, f0 minimum, standard deviations of values of f0 local peaks, mean of f0 peak rate, and jitter, and those associated with the second place (i.e. the target) were mean f0 peak bandwidth, mean f0 rates of rises, spectral emphasis, and HNR.

For all the line-ups, the melodic patterns appear as a relevant parameter to being used for prosodic-acoustic similarity purposes.

3.3. Perceptual analysis

The bi-dimensional plots for VQ and VD settings are shown in Figure 2. Dimensions D1 and D2 account for the transformation of the VPA scores into fitted distances in the MDS models. Data from RJ has two most chosen foil stimuli because both were equally chosen.

The normalized stress for the VQ analysis is 0.024, whereas for the VD analysis is 0.039, both considered acceptable values [12].

The analysis of the stimuli spatial arrangement sheds some light on listeners' false alarms, as well as on the possible reasons why some foils were not chosen.

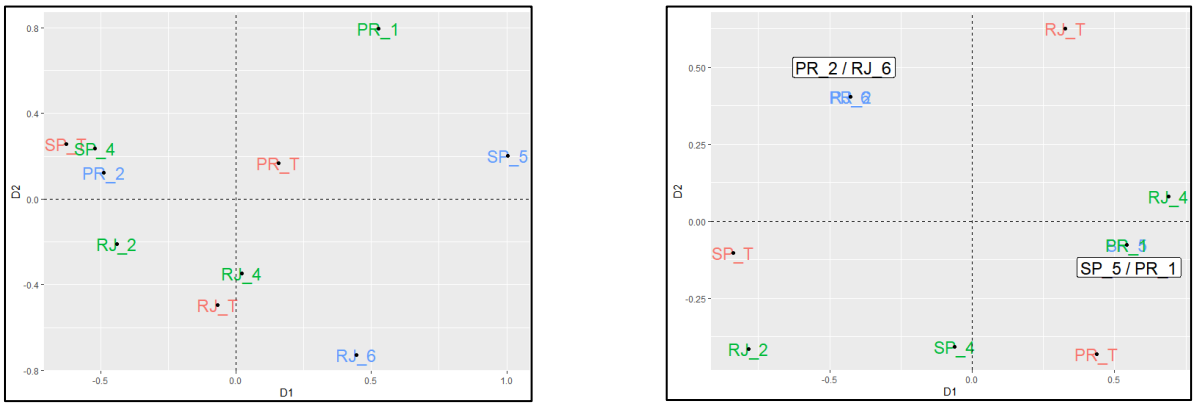

Figure 2: Bi-dimensional plots of the MDS analysis for the VQ (on the left side) and VD (on the right side) settings. Same line-up stimuli are identified by the acronyms of their states. The colours red, green, and blue stand for the target, the most chosen foil, and the non-chosen foil stimuli, respectively.

The comparison of the stimuli arrangement suggests that VQ may account for most of the false alarms in the SP ($d_{(SP\_T,SP\_4)} = 0.11$) and RJ ($d_{mean(RJ\_T, RJ\_2, RJ\_4)} = 0.47$) line-ups, whereas VD may account for most of the false alarms in the PR line-up ($d_{(PR\_T,PR\_1)} = 0.37$), since the target and the most chosen foil stimuli are closer to each other in the respective plots.

Regarding the non-chosen foil stimuli, it can be observed, in both plots, that the target and non-chosen foil stimuli pairs were placed on opposite sides of D1. The mean distances of these stimuli pairs suggest that VQ may also have been a relevant cue to perceived voice dissimilarity judgments in the SP line-up ($d_{(SP\_T,SP\_5)} = 1.63$), whereas VD may have been relevant to judging voice dissimilarity in the RJ ($d_{(RJ\_T,RJ\_6)} = 1.30$) and PR ($d_{(PR\_T,PR\_2)} = 1.20$) line-ups.

It is also worth mentioning that some stimuli were overlapped in the VD plot, which suggests that these speakers share similarities regarding VD settings. The "PR_2" and "RJ_6" stimuli, for example, share neutral settings for pitch and loudness variabilities, and for speech rate. The "SP_5" and "PR_1" stimuli, on the other hand, share non-neutral scalar degrees for minimized pitch range, low mean loudness, and slow rate.

4. CONCLUSION

Although it is very difficult (to not say almost impossible) to predict listeners' strategies when judging voice (dis)similarity, exploring prosodic parameters can give us some guidance. Therefore, the combination of acoustic and perceptual procedures proves to be complementary and necessary, as they delve into perception-production links.

In this study, the acoustic analysis showed that intonational, especially those regarding melody (e.g. mean f0 rates of rises and falls), and voice quality features are relevant for an overall voice similarity judgment. The perceptual analysis, on the other hand, shed some light on false alarms as well as on the non-chosen voices by analyzing voice quality and voice dynamics features separately. Moreover, it also added to the acoustic analysis by considering the analysis of speech rate features.

Special attention can be given to the PR_2 foil, which has not been chosen by any listener but was the first choice of the automatic analysis (i.e., the similarity function) for the PR line-up. These results highlight the importance of two aspects not considered in the acoustic analysis carried out here. The first one is the analysis of speaking tempo parameters, whose relevance to voice similarity had already been addressed by [3]. Speech rate probably also played a role among the voice dynamics settings in distinguishing this foil from the target, as observed in the VD plot. The other one is the analysis of segmental parameters, as mentioned by [4], which may signal sociolinguistic as well as idiosyncratic cues. In this specific case, the PR_2 foil produced an allophone of /t/ different from the target, and as this is a salient feature in some Brazilian dialects, it may have been one of the cues on which listeners relied to discard him. Therefore, speaking tempo as well as segmental parameters should be addressed in future studies.

Although voice line-ups are not part of the forensic casework in some countries, we believe they are useful to assess voice similarity in the forensic context. Furthermore, voice line-up analysis opens up possibilities to discuss why people confuse voices.

5. ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -

Brasil (CAPES) - Finance Code 001. Therefore, the authors acknowledge a grant from CAPES/PROCAD/Segurança Pública e Ciências Forenses #88887.516306/2020-00. The first author acknowledges a grant from CAPES #88887.804443/2023-00. The second author acknowledges a grant from PIPEq #21672, PUCSP.

6. REFERENCES

[1] McDougall, K. 2013. Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades. *International Journal of Speech, Language & the Law*, *20*(2).
[2] Broeders, A. P. A., Van Amelsvoort, A. G. 1999. Lineup construction for forensic earwitness identification: A practical approach. In *Proceedings of the 14th International Congress of Phonetic Sciences* 2, 1373-1376).
[3] Lindh, J. 2009. Perception of voice similarity and the results of a voice line-up. *FONETIK 2009*, 186.
[4] Lindh, J., Eriksson, A. 2010. Voice similarity-a comparison between judgements by human listeners and automatic voice comparison. *Working papers/Lund University, Department of Linguistics and Phonetics*, 54, 63-68.
[5] Nolan, F., McDougall, K., Hudson, T. 2011. Some Acoustic Correlates of Perceived (Dis) Similarity between Same-accent Voices. In *ICPhS* , 17, 1506-1509.
[6] Barbosa, P. 2021. *Prosody Descriptor Extractor* [Praat script]. Available: https://github.com/pabarbosa/prosodyscripts/tree/master/ProsodyDescriptorExtractor
[7] Traunmüller, H., Eriksson, A. 2000. Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, 107(6), 3438-3451.
[8] Mackenzie-Beck, J. 2007. Vocal profile analysis scheme: A user's manual. *Queen Margaret University College–QMUC, Speech Science Research Centre*, Edinburgh.
[9] Young, F. W. 1985. Multidimensional scaling. *Encyclopedia of Statistical Sciences*, University of North Carolina.
[10] Wickelmaier, F. 2003. An introduction to MDS. *Sound Quality Research Unit*, Aalborg University, Denmark 46.5, 1-26.
[11] de Leeuw, J., Mair, P. 2009. Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, 31(3), 1–30. Available: https://www.jstatsoft.org/v31/i03/
[12] Ashkiani, S. 2017. *Dimensionality Reduction and Data Visualization Using MDS-SMACOF package in R.* Available: https://rstudio-pubs-static.s3.amazonaws.com/246348_b31bca1e4be04bb395825dc6a00de364.html