Forensic Science International 270 (2017) 25–38

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics

Eugenia San Segundo[a,*], Athanasios Tsanas[b,c,d], Pedro Gómez-Vilda[e]

[a] *Department of Language and Linguistic Science, University of York, Heslington, York, YO10 5DD, UK*
[b] *Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK*
[c] *Wolfson Centre for Mathematical Biology, Mathematical Institute, University of Oxford, Oxford, UK*
[d] *Sleep and Circadian Neuroscience Institute, Nuffield Department of Medicine, University of Oxford, UK*
[e] *NeuVox Lab, Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain*

ARTICLE INFO

Article history:
Received 16 May 2016
Received in revised form 8 November 2016
Accepted 10 November 2016
Available online 17 November 2016

Keywords:
Forensic phonetics
Acoustic analysis
Perceptual assessment
Voice quality
Twins
Pause fillers

ABSTRACT

There is a growing consensus that hybrid approaches are necessary for successful speaker characterization in Forensic Speaker Comparison (FSC); hence this study explores the forensic potential of voice features combining *source* and *filter* characteristics. The former relate to the action of the vocal folds while the latter reflect the geometry of the speaker's vocal tract. This set of features have been extracted from pause fillers, which are long enough for robust feature estimation while spontaneous enough to be extracted from voice samples in real forensic casework. Speaker similarity was measured using standardized Euclidean Distances (ED) between pairs of speakers: 54 different-speaker (DS) comparisons, 54 same-speaker (SS) comparisons and 12 comparisons between monozygotic twins (MZ). Results revealed that the differences between DS and SS comparisons were significant in both high quality and telephone-filtered recordings, with no false rejections and limited false acceptances; this finding suggests that this set of voice features is highly speaker-dependent and therefore forensically useful. Mean ED for MZ pairs lies between the average ED for SS comparisons and DS comparisons, as expected according to the literature on twin voices. Specific cases of MZ speakers with very high ED (i.e. strong dissimilarity) are discussed in the context of sociophonetic and twin studies. A preliminary simplification of the Vocal Profile Analysis (VPA) Scheme is proposed, which enables the quantification of voice quality features in the perceptual assessment of speaker similarity, and allows for the calculation of perceptual–acoustic correlations. The adequacy of z-score normalization for this study is also discussed, as well as the relevance of heat maps for detecting the so-called *phantoms* in recent approaches to the biometric menagerie.

1. Introduction

The human vocal apparatus is a sophisticated system relying on the accurate synchronization of multiple organic structures (e.g. lungs, vocal folds, vocal tract) in order to produce speech. Researchers from diverse disciplines have approached this field from very different angles, and continue contributing to the understanding of this immensely complicated process. Traditionally, the structures involved in speech production have been separated into the systemic view of the source-filter model due to Gunnar Fant [1], where the laryngeal structures are credited for the production of phonation, and the supralaryngeal structures are credited for modifying phonation spectral contents dynamically. Although early works relied on the study of phonated speech as a whole, over the last years there is a growing consensus that hybrid approaches that take into account the source-filter distinction are needed for achieving more reliable techniques in Forensic Speaker Comparison [2]; hence this study undertakes the analysis of a set of voice features combining *source* and *filter* characteristics of the human voice.

State-of-the-art research on twins' voices [3,4] suggests that distinguishing this type of speakers poses a major challenge in speaker recognition because they are very similar. Extreme

* Corresponding author.
E-mail addresses: eugenia.sansegundo@york.ac.uk, eugeniasansegundo@gmail.com (E. San Segundo), tsanas@maths.ox.ac.uk (A. Tsanas), pedro@fi.upm.es (P. Gómez-Vilda).

http://dx.doi.org/10.1016/j.forsciint.2016.11.020

physical similarity also explains that other biometrics such as fingerprints [5] or palmprints [6] have been investigated in twins for identification purposes. In the case of forensic phonetics, including twins as participants in research experiments is of interest because these subjects may serve to assess how the results of pairwise comparisons – for the investigated voice characteristics – vary when highly similar speakers are considered (most often identical and fraternal twins but the variability of results can be observed considering also non-twin siblings or singletons). On the other hand, the relevance of twins is closely related with the search for robust voice characteristics for speaker discrimination, since a set of characteristics are considered robust for speaker comparison as far as they are maximally dependent on the speaker's genetic endowment and minimally influenced by learned factors, the latter favoring voice disguise or imitation. The predominance of genes over environment is thus linked to the two most important criteria for identifying characteristics for Forensic Speaker Comparison (FSC), namely that it should be as consistent as possible for each speaker, i.e. low intra-speaker variability, and that it should exhibit large variation amongst speakers, i.e. high inter-speaker variability [7,8]. Kinnunen and Li [9] refer to the same characteristics for an ideal Automatic Speaker Recognition (ASR) system.

Several acoustic parameters have been proposed to assess voice similarity in twins, the most common ones being fundamental frequency [10], formant patterns [11], or temporal characteristics [12], although ASR approaches are also common [13,14]. More recent investigations [3,15–17] have focused on the glottal analysis of twins, following a methodology that relies on the decoupling of the vocal tract from the glottal source estimates [18] and which allows the extraction of cepstral coefficients of the glottal source Power Spectral Density (PSD), singularities of the glottal source PSD, biomechanical estimates of vocal fold mass, tension and losses or time-based glottal source coefficients, among others. These features have the advantage of modeling the vocal folds and the vocal tract separately, which opens the possibility of independently studying source and filter information. The approaches in Refs. [3,15–17] present a clear advantage as far as the easy extraction of the speech material is concerned. In the cited studies, as well as in the present investigation, the glottal source features are extracted from naturally sustained vowels found in hesitated speech; also known as *fillers* or referred to as *disfluences* by other authors.

The main drawback for conducting more source-related studies in forensic phonetics in the past has been linked to the need for relatively long and stable vocalic sounds from which reliable values of distortion features like jitter and shimmer could be extracted. In clinical settings, these sounds are normally elicited upon asking the subject to sustain a vowel (typically [a]) for as long and steadily as possible [19,20]. This technique is unrealistic in a forensic context, but previous studies in Spanish suggest that [a] can be replaced by the use of naturally sustained pause fillers (typically [eː] in Spanish; [3]), as they are more forensically realistic while long enough for estimating a sufficient number of glottal cycles. This type of *disfluencies*, which are characteristic of spontaneous speech, have recently become a fruitful area of research interest. Künzel [21] already highlighted the consistency of speakers in their respective use of a personal variant of the hesitation sound, whether in relation to the addition of a bilabial nasal consonant or as regards the specific timbre of the vocalic component.

More recent studies have investigated formant values in filled pauses [22], or have focused on their duration and frequency of occurrence [23]. The extraction of voice quality features from fillers is less common [24]. The current study provides a new perspective to this type of disfluencies by analyzing 309 hybrid acoustic features to test their forensic potential in distinguishing same-speaker and different-speaker comparisons. This includes testing their robustness with very similar-sounding speakers, i.e. identical twins. In addition, this study explores novel methods for measuring (dis)similarities between subjects in pairwise comparisons, such as Euclidean Distances (ED). In twin studies, this type of statistical mapping has been recently used in Refs. [25,26]. Whereas both make use of ED, the former focuses on non-phonetic aspects (blood plasma lipidomics profiles), and only the latter is a phonetic study (a case study considering just one twin pair). In FSC in particular, French and colleagues [27] have explored ED to measure similarity between non-twin speaker pairs, including scores obtained from perceptual voice evaluations using the Vocal Profile Analysis (VPA) Scheme [28].

2. Materials and methods

This section presents the dataset used in the study and describes the methodology used to process the data. In the Section 2.2 we have distinguished between the acoustic analyses and the perceptual assessment of voices.

2.1. Data

We have used the phonetic corpus of Spanish male twins and siblings described in Refs. [3,29]. This comprises 54 speakers recruited ad hoc for the forensic phonetic investigation of twin and non-twin siblings in Spanish. To the best of our knowledge no other voice databases hitherto exist on twin voice research for the North-Central Peninsular Spanish variety. Although the database also includes dizygotic (DZ) twins and non-twin siblings, for this study we have only selected the available MZ twins (24 speakers) – all of the pairs having been raised together – and the group of unrelated speakers (12 speakers). The number of DZ twins was not enough to perform differential analysis; hence these samples were not considered.

Each speaker was recorded on two different occasions, separated by 2–4 weeks, in order to account for within-speaker variability. The two recording sessions took place in the Phonetics Laboratory of the *Consejo Superior de Investigaciones Científicas* (*CSIC*) in Madrid. The speakers were required to come in pairs for the voice recordings: with their co-twin in the case of MZ twins, and with a friend or work colleague in the case of unrelated speakers. This was aimed at attaining a comparable speaking style to what may be expected in conversations between twins, usually characterized by their spontaneity due to their close relationship. The age of the speakers of this database ranged between 18 and 52 years old (median: 28, interquartile range: 10). All participants were native speakers of North-Central Peninsular Spanish (see Ref. [30] for a description of this variety, also known as Standard Peninsular Spanish). A thorough questionnaire completed by all the participants served to assess health habits at the time of the recordings as well as to evaluate the degree of relationship closeness between pairs (only in the case of twins) by using Likert scales and typical questions used in previous phonetic studies on twins [11]. Besides, the zygosity of all the twins was checked; only for a MZ twin pair a DNA testing was necessary, which served to confirm that they were actually MZ twins.

Although the selected twin corpus included several speaking tasks, for this study we have only used the fifth speaking task: informal interview between each speaker and the first author of this investigation (the speaking styles exhibited by the participants were comparable to those found in forensic recordings). The interview lasted approximately 10 min and was carried out on the telephone, i.e. the researcher is at one end of the telephone and one member of each speaker pair at a time is at the other end of the telephone, in a different room. The recordings were made with high-quality but unobtrusive microphones (omnidirectional, condenser and flat-frequency-response microphones in an ear-set device). Forensically