Journal of Phonetics 75 (2019) 1–26

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Formant dynamics of Spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation

Eugenia San Segundo*, Junjie Yang

Department of Criminal Science and Technology, Shanxi Police College, No. 799, North-west Section, Qing Dong Road, Qingxu County, Taiyuan, Shanxi, China

ARTICLE INFO

Article history:
Received 30 August 2018
Received in revised form 28 March 2019
Accepted 1 April 2019
Available online 7 May 2019

Keywords:
Formant dynamics
Vocalic sequences
Diphthongs
Twins
Spanish
Curve fitting
Forensic

ABSTRACT

This study investigates the dynamic acoustic properties of 19 vocalic sequences of Standard Peninsular Spanish, showing their potential for forensic voice comparison. Parametric curves (polynomials and discrete cosine transform) were fitted to the formant trajectories of the 19 Spanish vocalic sequences of 54 male speakers, comprising monozygotic (MZ) and dizygotic (DZ) twin pairs, non-twin brothers and unrelated speakers. Using the curve-fitting estimated coefficients as input to a multivariate-kernel-density formula, cross-validated likelihood ratios were calculated to express the probability of obtaining the observed difference between two speech samples under the hypothesis that the samples were produced by the same speaker and under the hypothesis that they were produced by a different speaker. The results show that the best-performing system is one that fuses the 19 vocalic sequences with a geometric-mean fusion method. When challenging the system with related speakers, the results show that MZ twin pairs affect performance but, more importantly, that non-twin sibling pairs can deteriorate performance too. This suggests that more investigations are necessary into a range of similar-sounding speakers beyond MZ twins. Several nurture aspects are highlighted as explanatory factors for the strikingly high similarity of a specific non-twin sibling pair.

1. Introduction

This study investigates acoustic similarities in twin pairs and brothers using the formant trajectories of their vocalic sequences. The comparison of the speech patterns of identical and non-identical twins has attracted the attention of phoneticians in several areas of speech research. The recruitment of brothers and sisters is rarer (see Section 1.1), even though these are more easily found in the population than identical and non-identical twins. Investigations of twins typically comprise general descriptions of their degree of phonetic variability using a range of acoustic parameters or more applied investigations focusing on the clinical and forensic implications of twin research. In the clinical domain, investigations usually analyze infant speech and tend to focus on the search for heredity components of a vocal aspect (Forrai & Gordos, 1982; Matheny & Bruggemann, 1973). In forensic phonetics the research question revolves around which parameters are robust for speaker identification, although the term 'comparison' is preferred over identification (see Rose, 2003 or Morrison, 2009a). For this purpose, the recordings of same-sex adult twin pairs are compared to test the effect of related speakers on the performance of forensic systems based on one or several phonetic parameters.

All in all, the main reason why twin pairs are recruited as subjects for phonetic studies lies in the strong physical similarity existing between them, which is mostly due to their shared genetic information. Monozygotic (MZ) or identical twins develop from one zygote that splits and forms two embryos, so they typically share 100% of their genes[1] (Segal, 1990; Stromswold, 2006) while dizygotic (DZ) twins, also called non-identical or fraternal twins, develop from two separate eggs that are fertilized by two separate sperm cells (Abril et al., 2009).

*Corresponding author.
E-mail addresses: Eugenia@sxpc.edu.cn, eugeniasansegundo@gmail.com (E. San Segundo).

[1] Despite being frequently assumed to be so, a set of twins do not constitute a homogenous group as far as their genetic endowment is concerned. For instance, MZ twins can be monochorionic or dichorionic, depending on whether they share the same placenta or not. The fact that spontaneous mutations tend to occur more often in dichorionic MZ twins makes them more likely to differ genetically than monochorionic MZ twins (Stromswold, 2006).

https://doi.org/10.1016/j.wocn.2019.04.001
0095-4470/

They share an average of 50% of their genetic information. This makes the vocal tract anatomy of twins very similar and hence their voices. Another factor that contributes to their speech resemblance is their shared environment –in twins that have been raised together– which typically implies belonging to the same sociolinguistic group (e.g. family, school or even friends) and therefore, entails a certain degree of mutual vocal accommodation (San Segundo, 2014; Weirich, 2011; Zuo & Mok, 2015 describe this effect amongst twins). Besides nature and nurture factors, there is an often-neglected third factor accounting for twin (dis)similarities, i.e. epigenetics, which explains the alteration in the expression of specific genes caused by mechanisms other than changes in the underlying DNA sequence.[2]

Our current study seeks to investigate the forensic-phonetic potential of Spanish vocalic sequences –in terms of formant dynamics– in three main groups of related speaker pairs: MZ twins, DZ twins and non-twin siblings, all of them male. Full brothers are male siblings with the same father and the same mother, sharing 50% of their genetic information, like DZ twins. The phonetic-acoustic characteristics of these three groups of related speakers are studied in relation to a control population of unrelated speakers. This will serve as reference or background population in order to comply with the state-of-the-art methodology used in forensic studies to test system performance within a Bayesian framework (see Section 1.2 for an introduction to this methodology in a twin study, and Section 2.4.3 for more details).

1.1. Phonetic similarity of twins and non-twin siblings: forensic applications

Forensic phonetics is the application of phonetic knowledge to any type of legal issue, usually to tasks that arise out of a context of police work (Jessen, 2008). One of the most typical kinds of tasks involves the comparison of the voice of an offender (i.e. speech samples of an unknown speaker) with the voice of a suspect or several suspects (i.e. speech samples of known origin). This kind of task is referred to as Forensic Speaker Comparison or Forensic Voice Comparison (FVC henceforth). In FVC the recordings of the offender and the suspect/s can be compared using a wide variety of phonetic features, ranging from fundamental frequency and formant frequencies to articulation rate and voice quality as well as non-linguistic features (e.g. tongue clicking, audible breathing, throat clearing and laughter). Cambier-Langeveld (2007) and Gold and French (2011) document several international practices in FVC, including the most commonly analyzed phonetic parameters, comparison methods or reporting strategies. Ideally a good phonetic feature for FVC is one that varies as much as possible among speakers (high between-speaker variability) but remains as consistent as possible for each speaker (low within-speaker variability), as explained in Kinnunen and Li (2010) and Nolan (1983).

In this context, the importance of understanding how the speech patterns of twins vary with respect to other speaker populations can be derived. Research on this type of speakers allows testing the performance of a forensic-comparison system, since a robust system –and by extension, the parameters in which that system is based– should be able to distinguish even between very similar speakers. Since MZ twins are the most extreme cases of similarity in nature, they have traditionally been deemed to challenge FVC; in other words, this population can serve as the ultimate stress test to the performance of biometric systems. If these can accurately differentiate twins, then it is presumed that they will be able to easily separate unrelated individuals. The latter is the most common scenario in FVC, taking into account the low incidence of MZ twins worldwide. MZ twinning is thought to occur at a relatively constant rate of 3.5–4 per 1000 births across human populations (Bulmer, 1970; Hall, 2003; Smits & Monden, 2011).

San Segundo (2014) documented around forty studies delving into the voice similarities and differences of twin pairs, with a great majority of investigations focusing just on MZ twins. Over the intervening five years, recent investigations have appeared, notably in the field of forensic phonetics (da Costa Fernandes, 2018; Sabatier, Trester & Dawson, 2019; San Segundo & Künzel, 2015; San Segundo & Mompeán, 2017; San Segundo, Tsanas & Gómez-Vilda, 2017; Zuo & Mok, 2015). Due to the profusion of twin studies in Phonetics, a thorough literature review lies beyond our scope in this paper. We will briefly report on the main conclusions drawn from twin studies published so far, focusing on those with a forensic-phonetic component:

- From an auditory-perceptual point of view, most investigations conclude that twins' voices are highly confusable, making the task of twin voices' identification a difficult one, as has been shown in experiments where even twins were not able to distinguish their own voice from that of their co-twin (Gedda et al., 1960; Yarmey et al., 2001). Other studies suggest that listeners can tell twins apart by their voice above chance level (Decoster et al., 2000; Johnson & Azara, 2000). This makes researchers hypothesize that there must be some acoustical parameters which allow for speaker identification. Moreover, some studies found that MZ twin pairs can be as different between them as non-twin speakers usually are (Johnson & Azara, 2000), which has made many scientists wonder what proportion of inter-speaker variation is due to genetic factors and which to environmental reasons.
- The phonetic studies on twins which undertake an articulatory investigation are a minority and quite recent. The main findings from the scarce existing studies are limited to very concrete phonetic aspects like the phoneme contrast between /s/ and /ʃ/ in German investigated by Weirich (2011), who concluded that MZ are more similar than DZ twins in their articulation. The articulatory recordings of Weirich's investigation were carried out using a 2-D electromagnetic articulograph (EMA). For the articulatory measurements eight coils in total were attached to the subject's tongue, lips and jaw. Weirich (2011) complements her articulatory study with an investigation on perceived auditory similarity between twin pairs and with the analysis on certain acoustic correlates. Interestingly, she highlights the important role of lexical stress and coarticulation in her results.
- In terms of acoustic studies, many different acoustic parameters have been proposed to assess twins' (dis)similarities; most frequently, fundamental frequency (f0) and related parameters (Decoster et al., 2000; Debruyne et al., 2002), but also coarticulation patterns (Nolan & Oh, 1996; Whiteside & Rixon, 2004); Voice Onset Time (Ryalls et al. 2004); temporal parameters such as word and vowel durations (Whiteside & Rixon, 2001); voice quality

[2] See, for instance, Bruder et al. (2008) for more details about an investigation questioning the long-standing notion that MZ twins are essentially genetically identical.

features (San Segundo & Mompeán, 2017; Van Lierde et al., 2005; Weirich & Lancia, 2011) and formant patterns (Loakes, 2006; Zuo & Mok 2015). Among the main results obtained in these studies, some seem to point out to the difficulty or impossibility of discerning, for certain voice parameters, the influence of genetic factors from the influence of shared environment (e.g. f0: Debruyne et al., 2002). Based on their investigations of twins, other authors highlight the speaker-discriminant potential of particular speech features. Since the focus of this investigation is formant patterns, in next section we devote some space to the main conclusions of the few studies carrying out formant analyses in twins: Loakes (2006) and Zuo and Mok (2015).

As we mentioned above, forensic-phonetic investigations on non-twin siblings are not as common as twin studies. However, a few investigations exist which deserve some comments. Taking the data from Rose and Simmons (1996), Rose (2003) analyzed the F2 and F3 of five Broad Australian vowels extracted from stressed words in sentences pronounced by two brothers and their father. This study is based on Likelihood Ratios (LRs; see §1.2 for an introduction to this methodology in a twin study, and Section 2.4.3 for more details about LR calculation). The results of this investigation show LRs which support the same-speaker hypothesis for same-speaker comparisons (between direct and telephone speech for each speaker in non-contemporaneous recordings) and LRs which support the different-speaker hypothesis for the 12 possible different-speaker comparisons (considering both direct and telephone speech). These results are based on a combined LR (i.e. F2 and F3 of all five vowels together). However, some LRs derived from individual comparisons (i.e. taking individual vowels), ran counter to reality. Even though there was no attempt to take into account the possible correlations within the data (see *naïve Bayes* in §2.4.4) and despite the fact that the results do not distinguish between brother comparisons and son-father comparisons, this is to our knowledge the first LR-based investigation approaching sibling comparisons. The results are important because they point to the idea that "more is better" in LR-based FVC even (or maybe especially) for distinguishing between members of a family.

In a similar line to Rose (2003) –but without undertaking a LR-based approach– San Segundo (2010a) investigates formant frequencies (F1-F4) in five vowels of three Spanish-speaking adult brothers. Her results show that, in general, there are more significant differences in F3 and F4 than in F2 and F1, and that significant differences are more often found in between-speaker comparisons than in within-speaker comparisons (contemporaneous samples). However, there are interesting differences depending on the brother pair under comparison. While there are significant differences in the F4 of all five vowels between brothers Au. and C., it is F1 which yields significant differences between brothers An. and C. in all their vowels. Fewer significant differences were found between the pair Au and An. In addition, her results show that back vowels /u/ and /o/ –regardless of the formant considered– tend to obtain more significant differences in between-brother comparisons than in within-speaker comparisons (with respect to /a/, /e/ and /i/).

Other studies undertaking the analysis of brothers' speech, but less relevant for this investigation, are Charlet and Peral (2007), Feiser (2009), Feiser and Kleber (2012) or San Segundo (2014) (cf. San Segundo, 2014 for fuller descriptions). More recently, da Costa Fernandes (2018) has undertaken the investigation of certain speech characteristics of sisters together with female twins.

For a real case involving speech evidence and brothers in Australia, see Rose (2002). Two recent cases involving the arrest of MZ twins can be read in Calderwood (2015) and Himmelreich (2009). The former implied a robbery in Berlin but from the DNA evidence it could not be determine which brother took part in the crime, due to the shared genetic information between co-twins. The latter entailed the arrest of two MZ twins in France charged of six sexual assaults. Even though the victims claimed that the aggression took place by only one person, the police could not determine, on the basis of the DNA, which one of the two twins committed the crime. The reason set out by the police was that the DNA is the same for identical twins. The case was eventually solved as one brother confessed after he was given away by a stutter.

1.2. Formant dynamics and twins in previous forensic studies

When referring to the analysis of formant frequencies, the terminology "dynamic" vs. "static" is widespread in FVC thanks mainly to McDougall (2004, 2006), whose work on formant dynamics seems to be instigated by Nolan's important article (Nolan, 2002), where he suggests that "the imprint of an individual's speech mechanism (language, articulatory habits, and vocal tract anatomy combined) will be found to lie more in dynamic descriptions than in static descriptions" (Nolan 2002: 81). In other words, the phonetic realization of the formant trajectories in a VS would be strongly subject to the specific implementation of the acoustic targets by the speaker, naturally within his/her anatomical constraints. See Nolan (2002) for more details about the notion of 'phonetic target' and the idea that speakers present more acoustic similarities at the moments at which targets are achieved (e.g. the midpoint of formant frequencies) than at the transitional periods between targets. McDougall (2006) puts this in relation to other examples in human movement exhibiting highly individual differences. However, the suggestion that the study of the center of a vowel leaves much information unexplored for speaker characterization was already suggested by Goldstein (1976: 176): "The use of formant information in speaker identification systems have been limited almost exclusively to the measurement of formant frequencies inside a single window at the center of a vowel, leaving much of the formant structure unexplored (Sambur, 1975; Wolf, 1972)".

The few previous investigations on formant frequencies in twins (Loakes, 2006; Zuo & Mok, 2015) differ considerably in terms of methodology, number and type of twins investigated. Loakes (2006) follows a 'static' approach in the sense that she analyses the first four mean frequencies (F-pattern) of Australian English monophthongal vowels in four pairs of twins (three MZ twin pairs and one DZ twin pair). The research perspective undertaken by Zuo and Mok (2015) is 'dynamic' as they examine the first four formants of the diphthong /ua/ mea-

sured at each +10% step in eight pairs of Shanghainese-Mandarin bilingual MZ twins.

Despite following a traditional 'static' perspective, Loakes (2006) includes a quite novel[3] approach at that time, consisting in expressing the results of twin comparisons in likelihood-ratios (LRs). This methodological perspective –which is currently supported by most forensic phonetic experts– (see Section 2.4.3), presents the advantage of (1) taking into account within- and between- speaker variability and (2) assessing not only the similarity between speech samples but also their typicality with respect to an appropriate reference population. As suggested in Champod and Meuwly (2000), a Bayesian interpretation framework based on LRs is necessary to assess voice evidence in the field of FVC. Apart from concluding that F3 is the most speaker-specific formant frequency across diphthongs, the results of Loakes (2006) suggest that twins' speech is closer in F-pattern than the general population.

While Zuo and Mok (2015) do not compare the results of analyzing formant trajectories in MZ twins with those obtained by DZ twins or by a reference population, it is noteworthy that both studies (Loakes, 2006; Zuo & Mok, 2015) reach the same two conclusions: (1) that the degree of similarities in twin's formant frequencies is not uniform across twin pairs; and (2) that learned variation, or individual choice plays an important role in speech production and can account for the differences found between MZ twins. In both studies, the attitude towards being twins is highlighted as a contributing factor towards convergent formant patterns in twins. This is revealed in Loakes (2006) by noting that some MZ twin pairs made a conscious effort to sound different whereas DZ twins either were not so physically different from each other or used their vocal tracts in very similar ways. Likewise, Zuo and Mok (2015) found that the highly convergent formant patterns in certain twin pairs were most probably due to their strongest desire to signify their identity as twins – in comparison with other twins with a rather indifferent attitude towards having a twin. This phenomenon was particularly relevant to explain the finding that twins who were separated at birth were as similar, or even more similar, than some of the non-separated twins.

1.3. Spanish vocalic sequences

To the best of our knowledge, the present study represents the first investigation into formant dynamics in Standard Peninsular Spanish, in a larger population of twins and non-twins than previous studies, and using a Bayesian approach (see Section 1.4 for more details about how the current study differs from previous investigations). In particular, our study focuses on the whole set of vocalic sequences in Spanish. The term 'vocalic sequence' (henceforth VS) is used to refer both to the combination of vowel-vowel sequences as well as to the combination of glide-vowel sequences. In Spanish the first type of sequences are called hiatuses and the second type diphthongs (Aguilar, 1999), even though some terminological issues have been repeatedly highlighted regarding the phonological nature of diphthongs or the interpretation of glides (see Alarcos Llorach, 1965; Anderson, 1985; Aguilar, 1999; Hualde, 1991; Navarro Tomás, 1946; RAE, 2011). Hiatuses are also called heterosyllabic combinations (i.e. the elements making up the vocalic set belong to different syllables) while the label tautosyllabic combinations (i.e. belonging to the same syllable) is used to designate both diphthongs and triphthongs (RAE, 2011). Yet, these authors emphasize that the boundary between tautosyllabic and heterosyllabic combinations is not always clear.

Despite all these issues, the differentiation between hiatuses and diphthongs is considered "a genuine feature in Spanish" (Aguilar, 1999, p. 59). As explained by this researcher, "the fact that a sequence can be pronounced as a hiatus –i.e. in two separate syllables– or must be pronounced as a diphthong –that is, in a single syllable– is a lexical property: the acquisition of a new word implies the knowledge about its syllabification" (Aguilar, 1999, p. 59). Even though syllabicity is not defined in a precise way from a phonetic point of view (Aguilar, 1999), some acoustic cues have been highlighted for the hiatus-diphthong distinction, such as the formant transition rate (Borzone de Manrique, 1979; Quilis, 1981) as well as the onset duration, transition duration and offset duration (Borzone de Manrique, 1979). In RAE (2011: 337), the main differences between diphthongs and hiatuses lie on three characteristics: sequence duration, formant transitions and amplitude.

As regards duration, diphthongs would be shorter than their corresponding hiatuses. Although there is no agreement in this aspect, the transition between vowels would be shorter and quicker in hiatuses than in diphthongs. If we focus on amplitude, this parameter would be more similar between vowels in a diphthong than in a hiatus. Aguilar (1999) took a novel approach to find which acoustic cues –in the temporal and frequency domain– distinguish hiatuses from diphthongs. For that purpose, second-order polynomial equations were fitted to the F1 and F2 trajectories of 24 combinations of Spanish VS. The results of the study carried out by Aguilar (1999) show that hiatuses and diphthongs differ in both time and frequency domains, with hiatuses having a longer duration and a greater degree of curvature in the F2 trajectory than diphthongs. Besides, Aguilar (1999) found that there were differences between the two categories (hiatus and diphthong) depending on the communicative situation and that they behaved differently as far as phonetic reduction is concerned: "there is [...] an axis of reduction where a hiatus becomes a diphthong and a diphthong becomes a vowel" (Aguilar, 1999: 73).[4] For the purposes of our investigation, we are interested in testing whether similar types of curve parameterization methods, such as those used by Aguilar (1999), are useful for distinguishing speakers, rather than for a diphthong-hiatus differentiation.

Diphthongs have been traditionally classified in rising and falling diphthongs (Aguilar, 2010; Navarro Tomás, 1918; RAE, 2011). According to the description found in RAE

[3] We say 'quite novel' because –to the best of our knowledge– by the time the study of Loakes (2006) was published, only Rose (2006a) and Kinoshita and Osanai (2006) had explored the forensic discriminability of diphthongs from a LR-based perspective. Rose (2006a) examined five Australian diphthongs and concluded that diphthongs have considerable potential in FVC and needed to be researched more. In the same line, Kinoshita and Osanai (2006) examined the F2 slope in the glide of the Australian English diphthong /aɪ/ and found that this feature produced as good results as the F2 of the first and the second targets of that VS.

[4] According to Aguilar (1999: 73), "these results will argue in favour of the existence of a phonological structure shared by all the speaking styles, but with different phonetic manifestations in function of extralinguistic factors, such as the speaker's attention to his speech".

(2011: 332), in rising diphthongs, the vowel marked with the feature [+high] appears in the first position of the VS, while the vowel marked with the feature [−high] is in the second position. For the phonetic realization of these diphthongs (e.g. *miedo*, *justicia*, *tienda*), speech articulators move from a closure to an open position, making the second vowel in the sequence more salient. On the contrary, in falling diphthongs (e.g. *aula*, *boina*, *peine*), where the high vowel appears in second position, the speech articulators move from an open position to a closure. In these cases, the more salient vowel is the first one. Diphthongs can also consist of two different high vowels, like *ui*, as in *cuidas*. On numerous occasions, it has been said that the Spanish language favors diphthongization[5] , showing a clear tendency to avoid hiatuses (RAE, 2011:339). The RAE (2011: 333) adds that, furthermore, there is a preference for rising diphthongs in Spanish. This would be the reason why if two high vowels appear together, they form a rising diphthong, as in *buitre*, *ciudad* or *viudo*. Nevertheless, different factors may contribute to their realization as falling diphthongs, giving rise to pronunciation vacillations. This combination of two high vowels (group *iu* or *ui*) is particular prone to variation: while *buitre* or *cuita* are usually pronounced as diphthongs, the hiatus is preferred in words like *diurno* or *jesuita* (RAE, 2011: 337). According to Hualde and Chitoran (2003), exceptional hiatus (of the type iV) –leaving aside the cases with morphological or paradigmatic explanations– have a very restricted distribution. They occur only in stressed and immediately pretonic syllables (because these syllables tend to have greater duration than other syllables), but not further to the right.[6]

Concerning the elements of a diphthong, this type of VS are said to consist of a semivowel or semiconsonant and another vowel. The *i* and *u* vowels are pronounced as semivowels when they appear at the end of the diphthong, and as semiconsonants when they appear at the beginning. According to the International Phonetic Alphabet, the transcription of these elements is [i̯] and [u̯], whether they appear before or after the syllabic vowel (RAE, 2011: 333). This is the transcription convention adopted in this study.

For speaker comparison purposes, it is of special interest that both hiatus and diphthong pronunciations are sometimes allowed in Spanish, as we have briefly commented above. Considerable inter-speaker variation is therefore expected in most Spanish VS and in certain types in particular (see Section 2.2.1 Corpus design).[7] We deem that this fact could be useful for forensic purposes. Interestingly, since the pioneer investigations of Navarro Tomás (1918: 149) the creation of rules to regulate pronunciation vacillations in words such as *diana*, *crueldad* or *jesuita* has been considered pointless, given the several factors conditioning the two possible pronunciations and hence the speakers' freedom towards these VS (RAE, 2011: 337). The above-mentioned VS and, for example, others with the combination of a vowel with the feature [+high] and a vowel with the feature [−high] may be pronounced with hiatus or diphthong depending on several factors, not only geographic, sociolinguistic or stylistic, but also etymological or analogical.

A categorical division between hiatuses and diphthongs has not been attempted for the nearly 12 000 VS extracted for this investigation. Hualde and Prieto (2002) showed that native speakers of SPS do not necessarily agree in syllabification tasks when asked to divide words containing VS into syllables[8]. If a division between hiatus and diphthongs was to be made in future studies –in view of the results of this investigation– perhaps a newly-designed methodology would be necessary. If this division was to be based on a perceptual/subjective approach, at least more than one rater/observer would be required and interrater agreement should be properly evaluated. In order to perform the hiatus-diphthong division following objective/acoustic criteria, a large number of parameters –as detailed above– should be taken into account and properly analyzed. Hence the discussion of hiatus vs. diphthongs provided in this introduction has the aim of underpinning the choice of VS selected for our study as well as highlighting their issues and forensic potential, but not necessarily as a precursor to a phonetically-informed division within the results.

In any case, for clarification purposes –particularly for those not familiar with Spanish–, we provide two figures (Figs. 1 and 2). These show the spectrograms of a diphthong and a hiatus, pronounced by a male native speaker of SPS, aged 34 and instructed to pronounce the same word with a diphthong and with a hiatus in the VS /ua/ of the word *tatuaje* (tattoo).

1.4. The present study: Research questions

The present study seeks to fill the gap of previous twin investigations on formant frequencies inasmuch as we analyze these speech features from a dynamic point of view fitting the actual curves of VS. Besides, as a step forward from Zuo and Mok (2015), we approach the analysis of formant dynamics within the Bayesian framework for the evaluation of voice evidence, hence using LRs and a reference population. We also propose finer measurements for the characterization of the actual formant trajectories, namely two curve fitting methods: polynomial coefficients and Discrete Cosine Transform (DCT) coefficients. The combined use of curve fitting methods and LR-based results represent the state-of-the-art in phonetic studies of this sort (see, for instance, Morrison, 2009b; San Segundo, 2010b). In particular, the present study investigates the formant trajectories (F1, F2 and F3) of all the 19 VS of Standard Peninsular Spanish in a large population of 54

[5] The tendency to avoid hiatuses is especially prominent in fast speech (RAE, 2011: 349). This trend would explain many synaeresis and synalepha phenomena, being synaeresis the reduction to a single syllable of the vowels in a hiatus, taking place in a within-word context, while synalepha is the same phenomenon but occurring between words (RAE, 2011: 353).

[6] Hualde & Chitorán (2003) argue, therefore, that it is possible to explain, at least in part, the skewed distribution of exceptional hiatus in Spanish from general rhythmic patterns of the language. For other phonological accounts of the intricate relationship between VS and stress, see Colina (1999), Cabré & Prieto (2006) or Martínez-Paricio (2013).

[7] For instance, in some derivational words, there is a double stress pattern which affects the derivational suffixes and may therefore imply a double hiatus-diphthong pronunciation of some VS. As a case in point, the suffix *–iaco/-íaca ∼ -íaco/-iaca* admits both the proparoxytone and the paroxytone forms (RAE, 2011: 398).

[8] In the syllabification task designed by Hualde & Prieto (2002), only three of the six speakers recruited for the experiment showed total agreement with the hypothesized syllabification (i.e. the intuitions of the authors, who fully agreed on the syllabification) of 20 items containing the VS /ia/. This experiment is of special interest regarding Spanish stress because as the authors mention (p.220) "the majority of [*Spanish*] speakers do show complete agreement as to where the stress falls in all words, even if a small minority of Spanish speakers appear to be 'stress deaf'. This nearly universal agreement on intuitions is what makes the orthographic marking of stress practical in Spanish. For the hiatus/diphthong contrast, we still do not know this [. . .]".

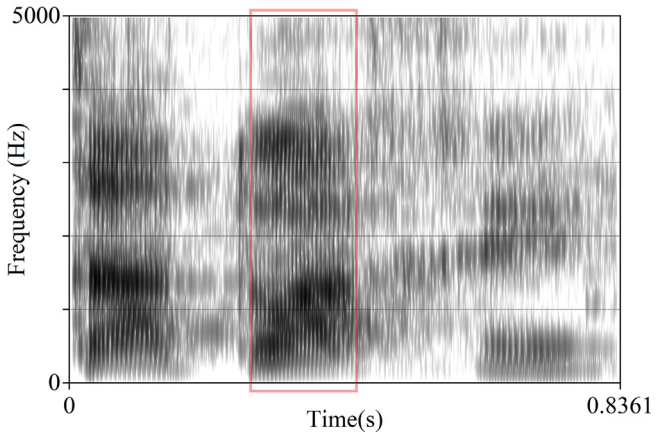

Fig. 1. Spectrogram showing the word [taˈtu̯axe] with a diphthongal pronunciation (zoomed in to show the sequence /ua/).

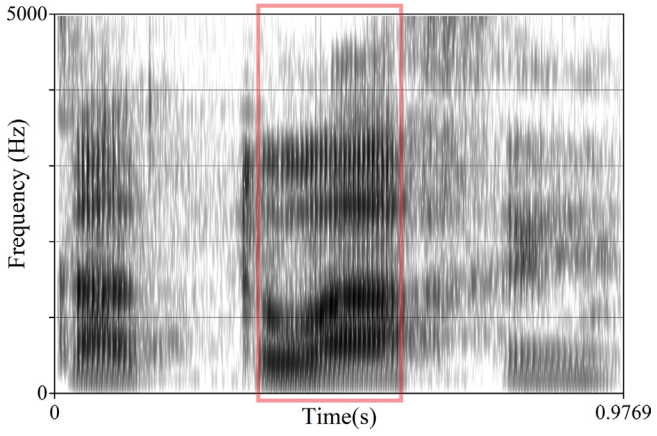

Fig. 2. Spectrogram showing the word [tatuˈaxe] with a hiatus pronunciation (zoomed in to show the sequence /ua/).

speakers, including identical twins, non-identical twins, brothers and unrelated speakers. For this reason, the present study also differs from previous investigations in a final aspect: the adding of both DZ twins and non-twin brothers as further examples of similar-soundingness. To the best of our knowledge, it is the first time that the combination of MZ twins, DZ twins, non-twin same-sex siblings and same-sex unrelated speakers are investigated together to explore the phonetic consequences of reducing between-speaker variation to a minimum (MZ twins) and to different similarity degrees in decreasing order (DZ twins, non-twin siblings and unrelated speakers).

The main objective of this study (RO-1) is to test the forensic-phonetic potential of formant trajectories extracted from Standard Peninsular Spanish VS. Other secondary research objectives are: (RO-2): to test the extent to which the fusion of 19 VS outperforms the individual VS for speaker comparison; and (RO-3) to test whether a certain procedure for curve fitting outperforms the other.

For the main objective, two research questions are raised: (RQ-1a) what is the overall performance of a forensic-comparison system based on VS formant trajectories?; and (RQ-1b): how do our different types of related speakers (MZ, DZ and B) affect the performance of this forensic-comparison system?

In relation to RO-2, it has to be borne in mind that adding more information does not necessarily mean that system performance is going to improve. As Rose, Osanai & Kinoshita (2003: 193) noted: “it is well known from automatic speaker and speech recognition that too much information can actually degrade performance”. However, recent investigations which follow a similar methodology to the one used in this investigation have shown that fusing the scores from individual speech features (e.g. vowels) tend to improve considerably on the individual features’ performance (e.g. González-Rodríguez et al., 2007; Rose, 2013). See Morrison (2013) for a discussion about the bias derived from increasing the number of parameters used by a forensic system, and how conversion of the scores to log likelihood ratios via logistic-regression calibration removes this bias.

2. Materials and methods

2.1. Subjects

The total number of speakers that participated in our study was 54, distributed in four different groups, as follows: (a) 12 MZ twin pairs, (b) 5 DZ twin pairs, (c) 4 pairs of full brothers, and (d) 12 unrelated speakers who are familiar to one another (i.e. 6 pairs of friends or work colleagues). All the speakers were asked to come in pairs to the recording sessions (see Section 2.3 for more details). In Section 1 we referred to the existence of two main types of twins, MZ and DZ, and we also mentioned that this investigation considers full brothers as well. Friends or work colleagues –the fourth speaker group– were recruited (1) in order to compare the results obtained by related speakers with those yielded by unrelated speakers, and (2) in order to create a reference population, whose relevance for LR-based studies will be explained in Section 2.4.3. The reason for recruiting friends and work colleagues was to achieve a speaking style similar, and thus comparable, to that found in the conversations between twins, usually characterized by their spontaneity due to a close long-term relationship.

The ages of the participants ranged between 18 and 52 years old (median age: 28.96). The age difference between the siblings in each pair varied between four and eleven years. In all cases they had an adult voice, neither presbiphonic nor adolescent. This study is limited to a single sex: male speakers. As concerns the dialectal aspects, the language variety spoken by all the subjects was North-Central Peninsular Spanish (see Hualde, 2005), also called Standard Peninsular Spanish (SPS henceforth).

Prior to participation in this study, the candidates had to fill in an online questionnaire aimed at assessing their suitability as regards the age and language criteria, besides gathering some other useful information like possible voice pathologies. The subjects who were finally selected for participation in this study had to fill in a more complete questionnaire at the day of the recording, which included questions about languages spoken, health habits, professional and leisure activities involving voice use and abuse, as well as relationship questions –in the case of twins– such as time typically spent together, communi-

cation habits, attitudes towards having a twin and experiences being perceptually misidentified.

2.2. Materials

2.2.1. Corpus design

The speech material used for this study was extracted from Task 2 of the Twin Corpus described in San Segundo (2013a) and San Segundo (2014). Task 2 was designed ad hoc to elicit the VS of SPS in a spontaneous context. The idea behind its design is that the speakers perform a collaborative task with the excuse of having received a fax copy (i.e. facsimile) in which not all the information is legible. This way speakers have to pronounce certain words that include the VS under investigation (see Section 2.2.3 for more details about the execution of the task).

The tool used for the word search was *BuFón: Buscador de Patrones Fonológicos* (English: searcher of phonological patterns), described fully in a manual available in Alves, Rico and Roca (2010). This online tool –with a search formalism based on the syntax of regular expressions– allows the user to insert the searched terms and it displays the results found with those characteristics in a corpus of texts extracted from the press and from dictionaries. The tool includes a phonological-search mode, which was very useful for the purposes of our study. For instance, as 'h' is not pronounced in SPS, the search term "ao" displayed both orthographic correspondences (e.g. *baobab* and *bacalao*) and phonological correspondences (e.g. *ahorrador* and *zanahoria*).

Concerning the search criteria of this tool, the following options were selected: "phonological" in 'search mode', and "press and proper names" in 'databases'. We preferred to consult only the press database and not the dictionary database since most of the words found in both databases were the same. Using both implied finding redundant information. Furthermore, only the press database contains details about word frequency and this is an aspect that we wanted to take into account for selecting the words which would eventually make up the fax sheets. Likewise, we opted for selecting the search option "proper names" because they were useful for the subsequent creation of the fax sheets (see Section 2.2.3). Besides, certain VS were found almost exclusively in surnames: e.g. "áe" (*Sáez, Herráez, Arráez, Peláez*).

As far as the search syntax concerns, it is worth noting that in the case of "ue" and "ui", it was necessary to specify that we did not want the program to display examples in which those sequences were preceded by "q" or "g", since in those cases "u" does not have a phonetic manifestation in SPS (e.g. *que* [ke]). The following syntax allowed us to avoid such examples: ¬[qg]ue, ¬[qg]ui.

In a first step, we searched for examples of the 20 VS of Spanish (*ae, ao, ea, eo, oa, oe, ou, ai, ei, oi, au, eu, ia, ie, io, ua, ue, uo, iu, ui*). The use of the *BuFón* tool served to discard the diphthong /ou/, as no words with the combination "ou" were found; only the compound word *estadounidense*. The rest of words were foreign loanwords (e.g. *glamour*, *country*, *boutique*) which have not been considered for this study since they do not have a pronunciation /ou/. These results agree with previous descriptions of this diphthong. For instance, Aguilar (2010) reports that the diphthong /ou/ has been considered rare in Spanish, as no Latin-origin words contain it. This VS was therefore discarded from our corpus.

As concerns the rest of vocal combinations, we distinguished between unstressed sequences (e.g. *israelí*) and stressed sequences. In these latter, we made a further distinction: stress in the first vowel (e.g. *Sáez*) or in the second one (e.g. *Rafael*). For each type, two examples were selected (see Table 1). Since there were 3 types of VS (unstressed, with the stress in the first vowel and with the stress in the second vowel), and 19 VS, the total number of words making up the corpus would be 114 (19 sequences × 3 types × 2 examples). However, eight words of the corpus contain two VS within the same word (*Beng**oe**ch**ea***, *p**oe**sí**a***, *c**ue**sti**ó**n*, *fisi**o**terap**eu**ta*, *di**e**ci**séi**s*, *c**eu**tí**e**s*, *ju**i**ci**o**so*) and there is also one compound (*jal**ea** r**eal***), considered one item with two VS. Therefore, the total number of words in our corpus is 106. The fact that some words contained two VS was very convenient in order to reduce the total number of words which should be produced by each speaker and so that the speaking task would not be very long and tedious.

Besides the 106 words making up the main corpus (see Table 1), we considered worthy adding 12 further words since their VS are expected to have particularly variable pronunciation between speakers (see Table 2).

In sum, the corpus is made up of 118 words: 106 containing the 19 types of VS of SPS (Table 1), plus 12 extra words which contain further examples of four VS in particular: *ue*, *ie*, *ua* and *ia*. The pronunciation of these was expected to trigger idiosyncratic between-speaker variation, since different pronunciation vacillations have been reported for them in the literature.

The three main criteria for the selection of the 118 words making up the corpus were: (1) frequency of occurrence (most frequent words were given preference, according to the occurrence percentage provided by the *BuFón* tool); (2) different within-word location of the VS; e.g. *espontáneo* (post-tonic position) and *leonés* (pre-tonic position); and (3) preference for consonantal contexts conveying less coarticulation. Following Marrero et al. (2008), we favored the following contexts: voiceless plosives [p,t,k], fricatives [s,f,z], rhotic [r] and affricate [tʃ], instead of nasals, voiced plosives and approximants.

2.2.2. Speech material

The speech material consisted of 11 773 phonetic units (i.e. VS). This number results from: 54 speakers × 2 recording sessions × 19 types of VS × 3 stress variations × 2 examples of each stress condition. The product should be 12 312 but some tokens had to be discarded for one of the following reasons: (1) non-modal phonation, like creak (in most cases) or whisper; or (2) overlap of the phonetic unit of interest with extraneous noises. The application of these exclusion criteria resulted in the selection of only homogenous tokens. This fact did not prevent at least one example per stress condition and type of VS being selected.

Prior to the analysis of formant dynamics (see Section 2.4. Measurements), the speech material required extraction and labelling. Firstly, the sound files of Task 2 (around 20–30 min) were cut into smaller files (10 min), which were then cut using the software *Sound File Cutter Upper* (Morrison, 2010b), allowing the cut-up of sound files and discarding the silent portions. This was found useful for an easy handling of

Table 1
Words containing the investigated vocalic sequence (VS).

VS	Words containing the VS: two examples per stress condition					
	Unstressed VS		Stress in the first vowel		Stress in the second vowel	
ae	*israelí*	*aeróbic*	*Herráez*	*Sáez*	*maestro*	*Rafael*
ao	*baobab*	*ahorrador*	*bacalao*	*Laos*	*Paola*	*zanahoria*
ea	*argéntea*	*bronceador*	*jalea (real)*[a]	*Bengoechea*[a]	*teatro*	*(jalea) real*[a]
eo	*espontáneo*	*leonés*	*boxeo*	*feo*	*león*	*gaseosa*
oa	*Joaquín*	*toallero*	*anchoa*	*Balboa*	*croata*	*almohada*
oe	*poesía*[a]	*Bengoechea*[a]	*aloe (vera)*	*Villarroel*[b]	*bohemio*	*soez*
ai	*faisán*	*vainilla*	*bonsái*	*káiser*	*bilbaíno*	*Países (Bajos)*
ei	*aceituna*	*voleibol*	*béisbol*	*dieciséis*[a]	*increíble*	*seísmo*
oi	*Moisés*	*boicot*	*hoy*	*Zoila*	*Eloísa*	*egoísta*
au	*auténtico*	*Paulina*	*Paula*	*flauta*	*Saúl*	*ataúd*
eu	*ceutíes*[a]	*mileurista*	*Ceuta*	*fisioterapeuta*[a]	*transeúnte*	*feúcho*
ia	*historia*	*asociación*	*poesía*[a]	*policía*	*estudiante*	*piano*
ie	*dieciséis*[a]	*ansiedad*	*ceutíes*[a]	*Díez*	*siete*	*viernes*
io	*fisioterapeuta*[a]	*funcionario*	*vacío*	*Ríos*	*juicioso*[a]	*cuestión*[a]
ua	*lengua*	*puntuación*	*cacatúa*	*ganzúa*	*donjuán*	*guapa*
ue	*cuestión*[a]	*pueril*	*bambúes*	*tabúes*	*sueco*	*cruel*
uo	*antiguo*	*mutuo*	*búho*	*flúor*	*Fructuoso*	*cuota*
iu	*ciudad*	*diurético*	*triunfo*[b]	*viudez*	*viuda*	*diurno*
ui	*ruiseñor*	*juicioso*[a]	*buitre*[b]	*fortuito*[b]	*suizo*	*genuino*

[a] Words with two VS.
[b] Words for which the combination of VS + stress condition did not exist. The following words have been used instead: *Villarroel* (stress in the second vowel of *oe*) since no words exist, besides *aloe*, with the stress in the first vowel; *triunfo* (stress in the second vowel of *iu*) and *viudez* (unstressed VS) since no word exists with the stress in the first vowel; *buitre* y *fortuito* (stress in the second vowel) since no word was found with the stress in the first vowel of *ui* (as explained in Section 2.3, if two high vowels appear together, they tend to form a rising diphthong).

Table 2
Extra words which make up the corpus and reasons for selecting them.

VS	Words	Reason for selecting them
ia	*mundial, oficial, viaje, confianza*	Words presenting pronunciation vacillation: they may be pronounced with hiatus or diphthong depending on several factors, not only geographic, sociolinguistic or stylistic, but also etymological or analogical.
ie ue	*hielo, hierro* *huevera, huevo, hueso*	Words beginning with *h-* plus *-ue* and *-ie* admit different degrees of plosive support before the VS.
ua	*tatuaje, suave, Atahualpa*	Words containing /ua/ are interesting for several reasons:(1) in *Atahualpa* high between-speaker variation is expected in the pronunciation of *h-* plus *-ua*, as it happens in *huevo* or *hueso*;(2) in words like *tatuaje* and *suave*, two trends have been traditionally observed: diphthong and hiatus pronunciations.

a series of short sound files in a later labelling step. For the labelling of the speech material, we used *SoundLabeller* (Morrison, 2012). As with other programs which allow the labelling of sound files, like the *TextGrid* function in *Praat* (Boersma & Weenink, 2012), this software displays the waveform and spectrogram of a sound file, enabling the user to mark the beginning and end of certain parts of the recording and to use labels for the selected fragments.

2.2.3. Execution of the speaking task

As explained above, Task 2 of the Twin Corpus was designed ad hoc to elicit VS in a spontaneous context, more accurately through an exchange of fax sheets by means of which speakers were forced to pronounce certain words that included the VS under investigation. Six different semantic contexts were devised for the six fax sheets that make up this speaking task. For instance, in the context of a Human Resources Department, one fax contains the names of several candidates for different jobs. Using this methodology, a range of words from Tables 1 and 2 were fitted into a role-playing context which required the exchange of information between speaker pairs. As a case in point, some of the words (VS in bold) that were elicited only in the first fax were: *P**au**la S**áe**z* (name of one of the job candidates), *fisi**o**terap**eu**ta* (occupation), *di**u**rno* (work shift), *mi**é**rcoles c**ua**tro de juni**o*** (date in which the interview would be held). Two copies per fax sheet were created: one per speaker. These two copies were not exactly the same. While some words could not be read properly in one copy, other words were illegible in the second copy. For the execution of this task, each speaker is in a different room and must follow these instructions: “You will find some fax sheets on the table. Their quality is not very good and some of the information on them is difficult to read. Your sibling/colleague has also received these fax sheets. Maybe his copies have a better quality than yours. Call him and ask him to give you the information that you cannot read properly in your fax sheet. Your brother will do the same with the information that is missing in his copy. Please provide him with the information that he needs”.

The Fax Task described here was adapted from the methodology described in Morrison, Rose and Zhang (2012) for the collection of forensic-phonetic databases. As in the Map Task (Anderson et al., 1991) –used by Aguilar (1999), among others– the aim of the Fax Task is to create a realistic context for the interaction between the participants; they have to gather information from each other so as to accomplish a common goal. A dummy fax was included at the beginning of the Fax Task in order to evaluate whether subjects had understood the task accurately and to detect ‘list effects’ in time. In that case, speakers were asked to avoid listing the missing words and instructed to insert them in a carrier sentence in a more natural way.

2.3. Recording procedure and data collection set-up

All the speakers were recorded twice, on two recording sessions taking place in different days, due to the importance of accounting for intra-speaker variability. The two sessions were separated by 2–4 weeks (mean: 22 days). For all the speakers and recording sessions we used the same recording material (microphones, soundcard and software) with the characteristics specified below. In addition, the recordings were always made by the same researcher –the first author– as a way to control that all the recordings were carried out using the same protocol.

Since the participants came to the recording sessions in pairs, two microphones were needed to record them at the same time. The microphones chosen for the recordings were two identical *Countryman E6i* earset microphones. These are omnidirectional condenser microphones especially suitable for this type of research since they are very small, thin and light, being thus unobtrusive. On the one hand, this helps the speaker to forget that he is being recorded, which is advisable in order to obtain spontaneous speech. On the other hand, since it is an earset device which is held close to the mouth, undesirable noise in the recordings is avoided. In addition, it ensures that the distance from the mouth to the microphone is always fixed. This microphone has a flat frequency response (20 Hz to 20 kHz), a sensitivity of 2.0 mV/Pascal, Equivalent Acoustic Noise 29 dBA SPL and Overload Sound Level 130 dB SPL. The microphones were connected to a soundcard through two long cables, each one to one channel. The soundcard was a *Cakewalk by Roland UA-25EX USB AudioCapture* with the following specifications selected for the recording: 44 100 Hz sample rate, 16 bits resolution, and mono channel. The software used for the recordings was *Adobe Audition CS5.5* and the telephone used for the communication between twins and between the researcher and the twins was a *Cisco IP Phone 7912 Series (Cisco Systems)*. Note that these telephones were only used by the participants to communicate with each other, as they were in separate rooms. To simulate real-condition telephone interceptions, the Twin Corpus also contains telephone-filtered recordings. However, the ones used in this study were the high-quality recordings.

Regarding the data collection set-up, the recordings took place always in the same setting. As previously explained, the speakers came in pairs to the Phonetics Laboratory of the National Research Council (Madrid). Here, they were first gathered together in the same room to receive the instructions for the different tasks. They were then separated in two quiet, almost identical rooms where the recordings took place. They were instructed not to provoke noise which could be undesirable for the proper recording of the acoustic signal, for example, moving the papers noisily or tapping the table.

2.4. Measurements

2.4.1. Acoustic analysis

The VS obtained following the steps described in the previous section were then analyzed with *FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories*, developed by Morrison and Nearey (2011). This software measures the formant trajectories of the specified segments using the formant tracking procedure outlined in Nearey, Assmann and Hillenbrand (2002). As specified in the software manual (Morrison & Nearey, 2011, p. 3), "the software measures formant trajectories using a range of parameters for linear-predictive-coding (LPC), runs some heuristics to attempt to identify the best track for each of the first three formants (F1, F2, F3), and presents the results to a human for checking". The formant trajectories are extracted using the algorithm described in Markel and Gray (1976) and the formants are tracked eight times using eight different cutoff values for F3 (2500–4000 Hz). As Fig. 3 shows, each of the eight formant-track sets are visually displayed per VS. These tracksets correspond to eight different F3-F4 cutoff values. Solid lines are used for the tracks from three-formants-below-the-cutoff, while the tracks from four-formants-below-the-cutoff appear as dotted lines. The F1-F3 tracks with thick lines are those determined to be the best on the basis of the heuristics (Morrison & Nearey, 2011). If users do not agree with the selected best track, they can choose other tracks. Fig. 4 shows the best formant-track set for one of the VS in our corpus.

2.4.2. Curve fitting

Once the F1-F3 trajectories of each VS were obtained, two different types of parametric curves were fitted to each trajectory: (1) polynomials of first, second and third order; and (2): first-through third-order Discrete Cosine Transforms (DCT). Curve fitting procedures are aimed at transforming a set of data points (the ones constituting the formant trajectories) into a small set of coefficients, thus performing data reduction.

The first type of curve fitting approximates the data points using polynomial functions of different degrees. The most basic polynomial function is the first-degree polynomial, which includes an offset or constant value (α_0) and a slope coefficient (α_1) which corresponds to the linear function (equation $y(x) = \alpha_0 + \alpha_1 x$). The second-degree polynomial function includes a quadratic term with a α_2 coefficient (equation $y(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$). For constructing the third-order polynomial functions, a cubic term with a α_3 coefficient is added (equation $y(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$).

The construction of a DCT function follows the same underlying idea that the polynomial curve fitting, but instead of using the linear, quadratic and cubic functions as basic elements, the DCT makes use of the sum of cosine functions with different amplitudes and frequencies as its building blocks or components (See Eqs. (1)–(3)).

$$y(x) = \frac{\alpha_0}{\sqrt{N}} + \frac{2\alpha_1}{\sqrt{N}} C_1 \tag{1}$$

$$y(x) = \frac{\alpha_0}{\sqrt{N}} + \frac{2\alpha_1}{\sqrt{N}} C_1 + \frac{2\alpha_2}{\sqrt{N}} C_2 \tag{2}$$

$$y(x) = \frac{\alpha_0}{\sqrt{N}} + \frac{2\alpha_1}{\sqrt{N}} C_1 + \frac{2\alpha_2}{\sqrt{N}} C_2 + \frac{2\alpha_3}{\sqrt{N}} C_3 \tag{3}$$

where N is the number of points in the original curve and C_k is the kth-degree DCT component.

For the rest of the investigation we decided to choose only the results coming from the best fitting parametric curve. For that purpose, we tested the goodness of fit of both types of parametric functions (and their three degrees) by means of linear correlation. As will be shown in the results (Section 3.1),

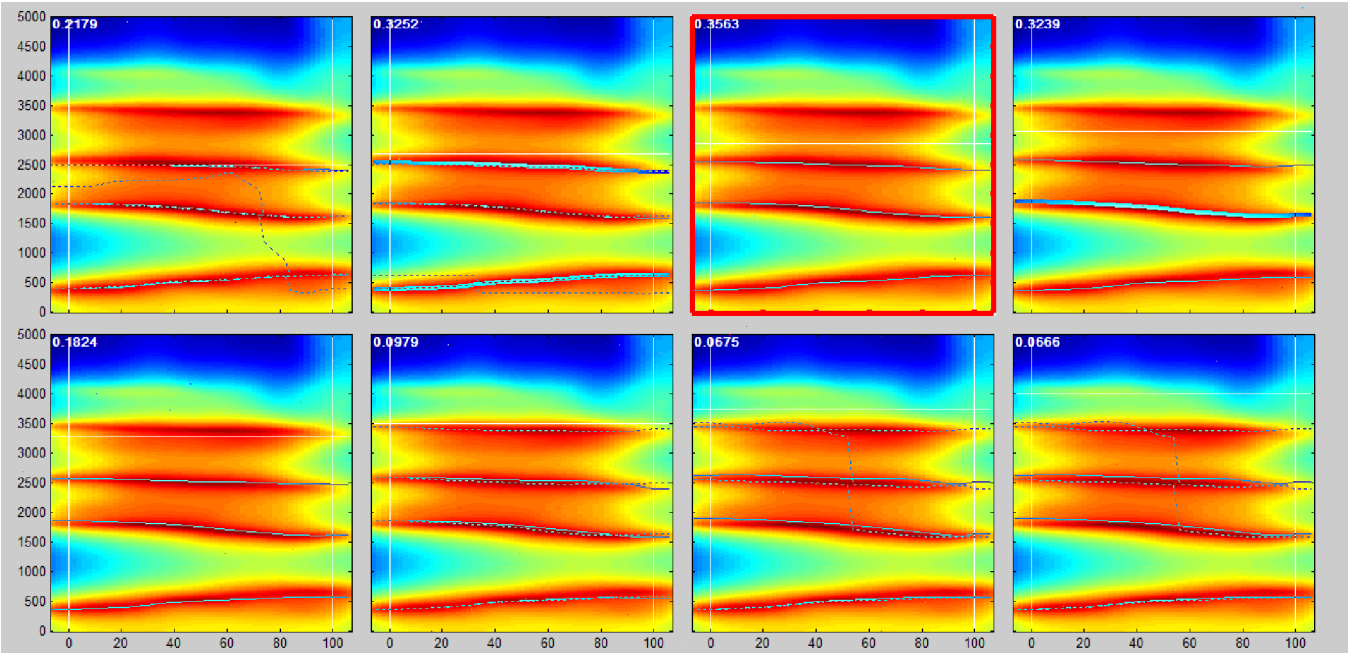

Fig. 3. Example of formant track selection. The different spectrograms show the eight possible formant-track sets for one of the diphthongs [ia] of speaker 04.

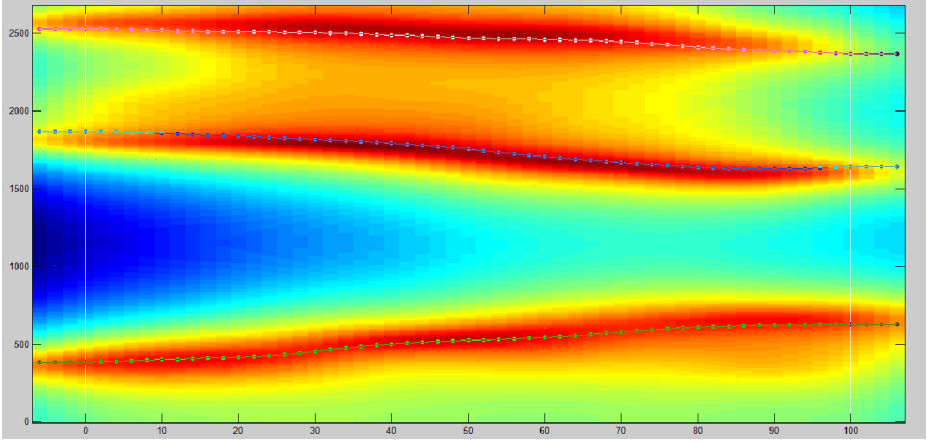

Fig. 4. Best formant-track set (F1-F2-F3) for one of the diphthongs [ia] of speaker 04.

both the cubic polynomial and the third-degree DCT outperformed their second-degree counterparts. In turn, the latter outperformed the first-degree functions. That is, a better approximation of the actual formant trajectory is achieved using this kind of curves. For the upcoming methodological stages we worked only with the results obtained from the cubic polynomial and third-degree DCT functions.

2.4.3. Likelihood-ratio calculation

For the LR calculation, we used the following input parameters: the coefficients obtained after approximating F2-F3 trajectories of the 19 VS by means of polynomial and DCT functions. We decided against including F1 following previous studies. For instance, Enzinger (2010) and Morrison (2009b) compared the performance of a system which fitted curves to the trajectories of F1, F2 and F3 with the performance of systems which only fitted curves to the trajectories of F2 and F3. Both of the above-mentioned studies concluded that the fused two-formant and three-formant systems yielded similar results, thus indicating that performance is not substantially deteriorated when F1 trajectories are not considered (Morrison, 2009). Besides, it is well known (Künzel, 2001) that the first formant is usually compromised by the telephone network passband (0.3–3.4 kHz) which affect telephone transmissions in real forensic casework.

In Section 1.2 we introduced the concept of LR, highlighting that they are the base of the Bayesian interpretation framework of voice evidence, and necessary to assess both the similarity between speech samples and their typicality with respect to an appropriate reference population. The LR is the numeric answer to the following question, posed by the court to the forensic expert: How much more likely are the observed differences between the known and questioned voice samples to occur under the hypothesis that the questioned sample has

the same origin as the known sample than under the hypothesis that it has a different origin? (Morrison, 2010a: 3052). The formula for the calculation of LR is provided in Eq. (4):

$$LR = \frac{p(E|H_{so})}{p(E|H_{do})} \quad (4)$$

where E is the evidence ("the measured differences between the samples of known and question origin" Morrison, 2010a, p. 3052); p(E|H) is the probability of E given H, H_{so} is the same-origin hypothesis, and H_{do} is the different-origin hypothesis. In the case of forensic voice comparison, H_{so} is typically represented as H_{ss} (same-speaker hypothesis) and H_{do} as H_{ds} (different-speaker hypothesis). For more detailed descriptions of this framework, see Berger, Robertson and Vignaux (2010), Champod and Meuwly (2000) and Ramos-Castro (2007).

For the LR calculation of the specific acoustic data of this study, we have used the Multivariate Kernel Density (MVKD) formula described in Aitken and Lucy (2004) and implemented by Morrison (2007). With this method it is possible to obtain LRs from continuous multivariate data. It was originally envisaged for the evaluation of trace evidence in form of glass fragments, but afterwards it has also proven useful for the forensic comparison of voice and speech evidence (Enzinger, 2010; Morrison & Kinoshita, 2008; Rose, Kinoshita, & Alderman, 2006).

The MVKD formula allows an evaluation of a) the similarity of two speech samples with respect to the intra-speaker variation and b) the typicality of the speech samples with respect to an estimate of the probability density of a reference population. In this formula, the within-speaker variance is estimated via a normal distribution, and the between-speaker population probability density is estimated via a kernel model (Morrison, 2009b). The multivariate data used in our investigation are the coefficients obtained after approximating the formant trajectories of the VS by means of polynomial and DCT functions.

A leave-one-out cross-validation procedure has been adopted in this study for the calculation of each LR. By means of this procedure, the background database consisted of data from all speakers except for the two speakers being compared each time. More specifically, according to this procedure, the data from each speaker's first session was compared:

(a) With the data from his own second session (which allow us to obtain non-contemporaneous intra-speaker comparisons).
(b) With the data from his brother's or speaking partner's second session. This allows us to obtain different-speaker comparisons of the following type: intra-pair MZ, DZ and B comparisons or just inter-speaker comparisons (for unrelated speakers, US).
(c) With the first session of all the other speakers in the background database (which gives further inter-speaker comparisons).

Thus, cross-validated LRs were calculated separately for each VS, represented by the curve-fitting coefficients of each of their F2-F3 formant trajectories. As explained above, we aimed at characterizing each speaker by both his F2-F3 trajectories together, having discarded F1. Formants are typically combined directly in the MVKD formula while VS needed a posteriori fusion (see Section 2.4.4).

2.4.4. Fusion techniques

After having obtained the LRs for each VS using the different types of parametric curves explained in Section 2.4.2, and having carried out the cross-validation procedure described in Section 2.4.3, the following step aimed at combining the results obtained per VS. This is done in order to improve the system performance (i.e. its potential for speaker individualization), according to state-of-the-art investigations on this topic (e.g. González-Rodríguez et al., 2007; Morrison 2009b). There are several methods for combining (summing or fusing) the results yielded by different systems. In our investigation, we have 19 different forensic-comparison-systems (as many as VS have been studied) that yield different scores (the way to call pre-fused LRs) and we aim at fusing them all in a single LR for each speaker comparison.

Two fusion techniques were implemented: (1) Naïve Bayes with a posteriori geometric mean calculation and (2) logistic regression. The first procedure assumes statistical independence of the scores (systems) to be combined, while the second one does not, and therefore needs some calibration.

In a first step, we combined the scores obtained from the 19 systems (one per VS) by simply multiplying them together. This procedure is called *Naïve Bayes* (also Idiot's Bayes or Independence Bayes; see Rose, 2006b) and it assumes that the variables are independent, i.e. they are not correlated. Therefore, the value of the combined LR (LR_c) will be calculated as shown in Eq. (5):

$$LR_c = Score_1 \times Score_2 \times Score_3 \times \cdots Score_{19} \quad (5)$$

Yet, in order to avoid an overconfidence of the *LRc* obtained, in a further step we proceeded to calculate the 19th root of the product, i.e. obtaining the geometric mean. Assuming statistical independence where there is actually correlation between variables, naïve Bayes fusion tends to yield overestimated LRs. Therefore, the calculation of the geometric mean of all the 19 scores instead of the simple product is recommended to compensate this overconfidence in the LRs (See Eq. (6)).

$$LR_c = \sqrt[19]{Score_1 \times Score_2 \times Score_3 \times \cdots Score_{19}} \quad (6)$$

The second type of score fusion is called logistic regression, a well-known statistical classification model (Hastie, Tibshirani & Friedman, 2009). In its forensic application, the use of logistic regression implies not only fusion but also calibration (e.g. Brümmer et al., 2007; Brümmer and du Preez, 2006; González-Rodríguez et al., 2007; Morrison & Kinoshita, 2008; Pigeon, Druyts & Verlinde, 2000; Ramos-Castro, 2007; van Leeuwen & Brümmer, 2007; in Morrison, 2010a). On the one hand, calibration is the process of designing and optimizing the transformation from the raw scores calculated by different systems into LRs in such a way that a cost function is minimized. On the other hand, fusion converts multiple sets of scores into LRs. In any case, what scores do is "quantifying the degree of similarity of pairs of samples while also taking account of their typicality" (Morrison, 2010a, p. 3061). These scores, or uncalibrated LRs, do not have an absolute meaning by themselves. However, it is the LR value after calibration what represents the weight of the evidence. For the application of logistic regression, it is necessary to have some training data:

scores from comparisons where it is known whether they are same-speaker comparisons or different-speaker comparisons.

For carrying out calibration and fusion we used the logistic regression functions in the *FoCal Toolkit* (Brümmer, 2005), both for the training part and the fusion part. In relation to the training stage, and considering the size of our database (not as large as the ones normally used in statistical studies due to the inherent limitations of twin studies) we needed to ensure that the training population did not include any of the subjects under test. For that purpose we trained a different model (and thus obtained different weights) for each of the comparisons carried out by the 19 MVKD systems. The trained model did not include any of the two speakers being compared, in order to fulfill an honesty criterion, namely, that the scores used for training must be different from the scores to be fused.

For more details about logistic-regression calibration and fusion at a practical conceptual level and minimal mathematical complexity, see the tutorial of Morrison (2013).

2.4.5. Accuracy assessment

Assessing the output accuracy of a forensic-comparison system is a very relevant aspect in forensic sciences. Several metrics and graphs have therefore been developed to evaluate such accuracy. For this study we have used the log-likelihood-ratio cost (C_{llr}) and Tippett plots.

The C_{llr} was originally envisaged for its use in automatic speaker recognition (Brümmer & du Preez, 2006; van Leeuwen & Brümmer, 2007) but has also been applied in forensic-comparison studies based in traditional acoustic parameters (e.g. González-Rodríguez et al., 2007; Morrison & Kinoshita, 2008). This measure is defined in Eq. (7):

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{H_p}}\sum_{i=1}^{N_{H_p}}\log_2\left(1+\frac{1}{LR_i}\right)+\frac{1}{N_{H_d}}\sum_{j=1}^{N_{H_d}}\log_2\left(1+LR_j\right)\right) \quad (7)$$

where N_{Hp} is the total number of LRs for the *Hp* (hypothesis of the prosecution) and N_{Hd} is the total number of LRs for the *Hd* (hypothesis of the defense). The LRs for the *Hp* are referred as LR_i and the LRs for the *Hd* are called LR_j. In a typical forensic situation, *Hp* equals H_{ss}, i.e. "the offender and the suspect samples are from the same origin (same speaker)" while *Hd* equals H_{ds}, i.e. "the offender and the suspect are different speakers". The C_{llr} will depend on these hypotheses.

According to the equation above, the lower the C_{llr}, the more accurate the performance of the system. This measure can be used to compare several systems which are based on the same set of data. For instance, we have compared for our study the performance of 19 systems, one per VS. On the assumption that target comparisons (LR_i) should yield high LR values and non-target comparisons (LR_j) should yield low LR values for a forensic system to perform optimally, any deviation from this ideal situation is punished, with highly misleading LRs being charged heavier penalty (i.e. higher C_{llr} values) and vice versa (cf. González-Rodríguez et al., 2007). So, for every comparison system, large positive LLR (log-likelihood ratios) for same-speaker comparisons and large negative LLR for different-speaker comparison are assigned very low C_{llr}. In contrast, as specified in Morrison (2010a), these C_{llr} values get higher and higher as the LLRs become more negative and provide stronger contrary-to-fact support for the different-speaker hypothesis. Since LLRs close to zero do not provide a strong support for either H_{ss} or H_{ds} they are assigned moderate C_{llr} values.

Normally, not only the single measure C_{llr} is provided but also the so-called $C_{llr\ min}$, which represents the C_{llr} obtained for a system without calibration errors. The difference between C_{llr} and $C_{llr\ min}$, known as $C_{llr\ cal}$, yields a numeric value which represents the calibration loss of the system.

Tippett plots represent another method for evaluating the performance of a forensic-comparison system but, as compared with the single measuring value of the C_{llr}, Tippett plots are graphical representations where more information can be found about the output of a LR-based comparison system. In this type of graph (proposed by Evett and Buckleton (1996) in the field of DNA analysis), two curves are displayed, each one representing the probability for one of the competing hypothesis. Usually the hypothesis of the prosecution is that the offender and the suspect samples come from the same speaker, while the hypothesis of the defense (H_d) is simply that they belong to different speakers. However, for the speaker types that we are testing (MZ, DZ, B or US), we will draw Tippett plots based on a different H_d (i.e. that the speech samples belong, not to the same speaker, but to his MZ, DZ, B or US, correspondingly).

2.4.6. Summary of methodological steps

Since the methodological steps followed for this investigation are manifold, Fig. 5 has been designed with the aim of summarizing in a diagram the different stages carried out for the complete analysis of formant trajectories, from the speech material extraction to the accuracy evaluation of the forensic-comparison systems. These stages are fully described in Sections 2.4.1 to 2.4.5.

3. Results

3.1. Curve fitting: best correlation values

Several correlation tests were calculated between the coefficients of the original formant (F2 and F3) trajectories and their fitted curves, using both (first-through-third order) polynomial and DCT coefficients. This method was used to calculate the goodness of fit of each function. The results showed that second- and third-degree functions (regardless of whether it is polynomial or DCT curves), presented much higher correlation values than the first-order functions. Therefore, Table 3 shows only the correlation results for second- and third-degree curve fitting.

Table 3 shows that the approximation of F2 trajectories (R values in bold) is always better than the fitting of F3. This occurs for all VS and regardless of the type of parametric curve. In addition, the third-degree functions (R values in italics) outperform their second-degree counterparts, again regardless of the VS and the type of parametric curve. Therefore, the best correlation values always correspond to third-degree functions and F2 (bold and italics).

Regarding which VS are worst and best fitted, /uo/ seems to be the VS where a worst fitting is achieved, comparatively. This occurs for all types of curve fitting and degrees, and both for F2 and F3. For example, its R values for F3 curve fitting range between 0.8705 (DCT) and 0.8878 (polynomial function). In

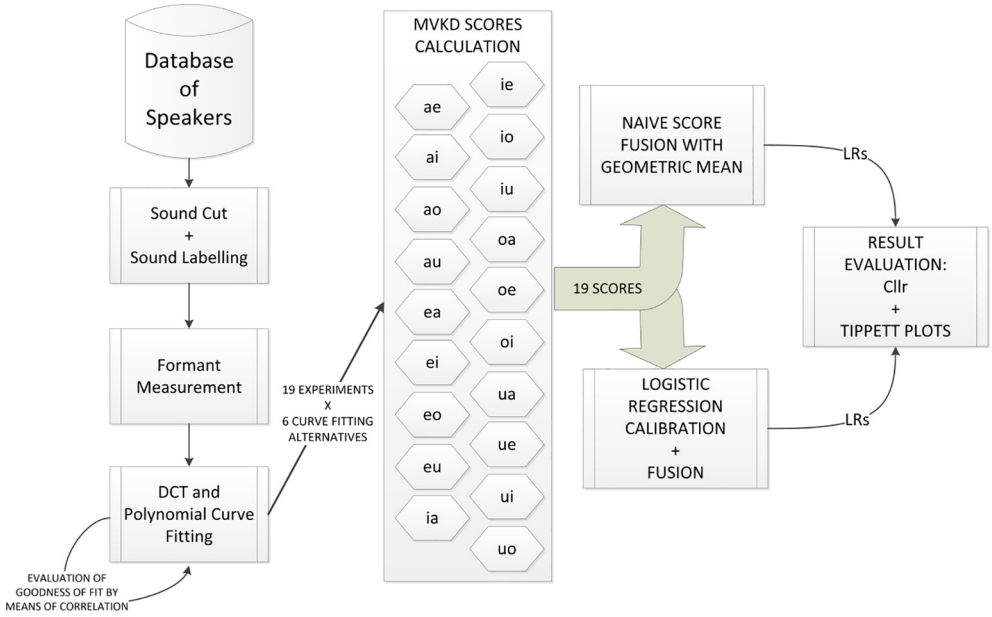

Fig. 5. Diagram showing the different stages carried out for the analysis of formant trajectories in Spanish VS. With '19 experiments' we mean that we calculated MKVD scores 19 times, one per VS. With '6 curve fitting alternatives' we mean three coefficient orders per curve fitting method (polynomial and DCT).

Table 3
Correlation coefficients between the original formant (F2 and F3) trajectory and their fitted curves (polynomial and DCT).

		Type of curve fitting			
		Polynomial		DCT	
Vocalic sequence	Formant trajectory	Quadratic	Cubic	2nd-degree	3rd-degree
/ae/	F2	**0.9800**[a]	***0.9923***[b]	**0.9844**	***0.9915***
	F3	0.8486	*0.9058*	0.8588	*0.9132*
/ai/	F2	**0.9816**	***0.9941***	**0.9908**	***0.9952***
	F3	0.9112	*0.9531*	0.9250	*0.9607*
/ao/	F2	**0.9378**	***0.9851***	**0.9511**	***0.9809***
	F3	0.8401	*0.9023*	0.8360	*0.8939*
/au/	F2	**0.9273**	***0.9845***	**0.9451**	***0.9799***
	F3	0.8138	*0.9156*	0.8164	*0.9075*
/ea/	F2	**0.9673**	***0.9897***	**0.9790**	***0.9904***
	F3	0.8205	*0.8907*	0.8383	*0.9050*
/ei/	F2	**0.9658**	***0.9860***	**0.9619**	***0.9765***
	F3	0.8765	*0.9380*	0.8832	*0.9380*
/eo/	F2	**0.9600**	***0.9905***	**0.9790**	***0.9932***
	F3	0.8496	*0.9194*	0.8654	*0.9302*
/eu/	F2	**0.9350**	***0.9843***	**0.9638**	***0.9896***
	F3	0.8369	*0.9023*	0.8548	*0.9201*
/ia/	F2	**0.9743**	***0.9911***	**0.9844**	***0.9915***
	F3	0.8555	*0.9295*	0.8713	*0.9327*
/ie/	F2	**0.9710**	***0.9886***	**0.9740**	***0.9850***
	F3	0.9116	*0.9599*	0.9228	*0.9598*
/io/	F2	**0.9714**	***0.9919***	**0.9845**	***0.9941***
	F3	0.8849	*0.9430*	0.9029	*0.9527*
/iu/	F2	**0.9551**	***0.9894***	**0.9790**	***0.9944***
	F3	0.8837	*0.9334*	0.9028	*0.9524*
/oa/	F2	**0.9684**	***0.9885***	**0.9714**	***0.9842***
	F3	0.8494	*0.9151*	0.8447	*0.9095*
/oe/	F2	**0.9726**	***0.9940***	**0.9880**	***0.9959***
	F3	0.8284	*0.9091*	0.8422	*0.9239*
/oi/	F2	**0.9698**	***0.9900***	**0.9853**	***0.9937***
	F3	0.8539	*0.9134*	0.8704	*0.9351*
/ua/	F2	**0.9686**	***0.9898***	**0.9756**	***0.9873***
	F3	0.8375	*0.9119*	0.8396	*0.9113*
/ue/	F2	**0.9819**	***0.9940***	**0.9904**	***0.9951***
	F3	0.8186	*0.9195*	0.8202	*0.9150*
/ui/	F2	**0.9689**	***0.9892***	**0.9837**	***0.9928***
	F3	0.8907	*0.9444*	0.8989	*0.9543*
/uo/	F2	**0.9310**	***0.9707***	**0.9289**	***0.9591***
	F3	0.8008	*0.8878*	0.7897	*0.8705*

[a] We highlight in bold the R values for F2, which are always larger than for F3, regardless of the VS or the curve fitting procedure.
[b] We highlight in italics the R values obtained with cubic polynomials and third-degree DCT functions, which are larger than with their quadratic and second-order counterparts, across VS types and formants.

contrast, there is not a single VS which always obtains the maximum R value. It depends on the parametric function and formant considered. In general, we can observe that /ai/, /ie/, /ue/ and /oe/ tend to obtain the highest correlation, with R values around 0.99, especially in F2 fitting. As explained in Section 1.3, results are not partitioned in hiatuses and diphthongs.

3.2. Unfused and fused results: system performance

We explained in Section 2.4.4 that after obtaining the LR-results for each individual VS (before calibration, these are called scores), we would proceed to their fusion in order to improve their forensic performance. Fig. 6a-b include information about the accuracy of the individual systems based on a single VS and the fused 19-VS system. Having found that the third-order functions –both for polynomial and DCT functions– fitted the real formant trajectories better than their second-degree counterparts (see Section 3.1), we considered only the fusion of scores derived from those functions (Poly3 and DCT3 henceforth).

Following the method proposed in previous forensic studies on twins (e.g. Künzel, 2011; San Segundo & Künzel, 2015), we first present the performance results of our system without separating our speaker population per type of speaker (MZ, DZ, etc.) and then we provide the results per speaker type to see how each of them affects the performance of the system. Next section (Section 3.3) tackles the question of whether there is higher intra-pair similarity in the formant dynamics of MZ twins than in other speaker comparisons (DZ, B or US).

As explained in Section 2.4.4, the first procedure that we carried out to combine the multiple scores resulting from the 19 different systems (one per VS) consisted in simply multiplying them together *à la naïve* Bayes. To the product of the multiplication, we further calculated the 19th root with the aim of obtaining the geometric mean. The purpose of this was to compensate the overconfidence expected in the LRs obtained with

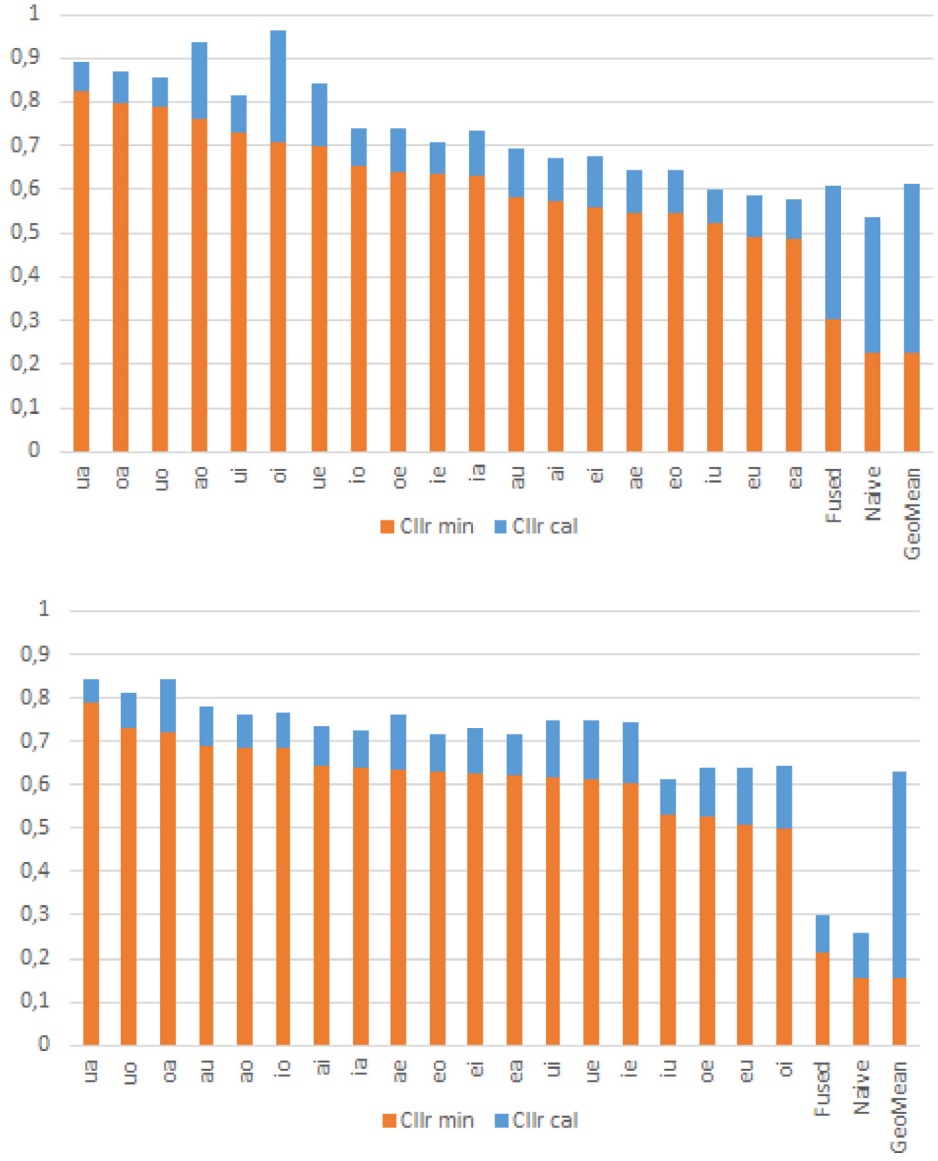

Fig. 6. a-b. C_{llr} plots for the unfused and fused 19-vocalic-sequence systems. $C_{llr\ min}$ represents the C_{llr} obtained by each system after calibration (i.e. without calibration errors). $C_{llr\ cal}$ is the difference between the C_{llr} and the $C_{llr\ min}$ (i.e. the calibration loss of the system). Fig. 6a (top): results for the third-order polynomial curve fitting method; Fig. 6b (bottom): results for the DCT curve fitting method.

the naïve approach, which omits any consideration about the structure of the data to be fused, i.e. assumes statistic independence of each of the 19 systems to be fused.

We explained in Section 2.4.5 how the cost function C_{llr} is used to evaluate the performance of speaker comparison systems (the lower the C_{llr}, the more accurate the performance of the system). Both Fig. 6a and b show that the use of fusion techniques improves the performance of the forensic-comparison systems. While the C_{llr} values obtained when considering each VS independently are as high as 0.82 after calibration ($C_{llr\ \mathrm{min}}$), yielded by /ua/ when considering Poly3 (Fig. 6-a), in a combined 19-V-S system this $C_{llr\ \mathrm{min}}$ value drops considerably: till 0.30 when using logistic-regression fusion and till 0.22 when combining results with both the naïve and the geometric-mean methods. When considering DCT3 (Fig. 6-b), similarly high C_{llr} values are obtained with any VS in isolation, /ua/ being the highest again with 0.78 in the $C_{llr\ \mathrm{min}}$. The application of fusion techniques entails a drop in the function cost, with a 0.21 $C_{llr\ min}$, obtained with the logistic-regression fusion, and a 0.15 $C_{llr\ min}$ obtained with the naïve and the geometric-mean sum. The implication is that the best-performing forensic system is one that characterizes each speaker by all their VS, not individual ones. Even though this result was somehow expected, we noted in the introduction that more information is not necessarily better for system performance. All in all, our objective was finding out to which extent a fused system has better performance than the best individual VS. As said before, the $C_{llr\ min}$ value drops considerably in a combined 19-VS system. It seems therefore useful to consider as many VS as possible for forensic cases. It is true, however, that extracting as many as 19 could be too time-consuming –if it is ever possible to find an example of each one in real forensic recordings, usually characterized by their short duration. Different combinations of VS should be investigated to find a system that achieves maximum performance with the minimum number of VS.

When evaluating the logistic-regression fused scores of the system, we observe that they do not offer a better performance in terms of $C_{llr\ min}$, compared to the geometric mean or the naïve combination. Normally, this means that the training is not converging correctly. In order to confirm this, we tested a modified version of *FoCal*'s training function, which includes a scaling parameter called *lambda* that makes LRs lower, but it is associated with a more robust convergence of the training step. Lambda is a regularization factor used to evaluate the convergence of the logistic regression fusion. As it did not seem clear why the logistic-regression fusion yielded worse results than the other methods, we hypothesized that this could be due to a lack of enough training data, whose origin would be in the small database used. The goal of lambda is therefore to mitigate the effect of the lack of data for the training of the logistic regression, although at the cost of yielding LRs which are more moderate (under-confident), i.e. less strong than they should.

After evaluating the C_{llr} values obtained with the above-mentioned modified training function, we observed that the lower the lambda, the lower the $C_{llr\ min}$. This helped us confirm that our logistic-regression model was diverging because our database was lacking a larger set of data for the training, and even if we did an "honest" training (i.e. the scores used for training were different from the scores to be fused), it was not yielding accurate results. Therefore, the validity of this kind of approach (i.e. logistic regression) needs to be questioned in this kind of situations, i.e. when a small database is being used. Furthermore, it should be noted that a modification of the training function provided by *FoCal* has been made with the aim of using a more exigent threshold. Despite this, the regression model did not converge either. In other words, the results presented in Fig. 6a-b using the standard function in *FoCal* have been obtained using a convergence threshold of 10–12 instead of 10–5, which is the default value in *FoCal*. The meaning of this threshold is as follows. After finishing each of the iterations, a calculation is made of the difference between the new weights obtained in such iteration and the weights which have been obtained in the previous iteration. If the difference is smaller than the established threshold, the training phase is considered finished with the last obtained weights. This reduction in the convergence threshold implies that for our study we have been more exigent in relation to convergence. Despite this, convergence of the model has not been attained optimally, as has been shown with the use of the correction factor lambda.

All in all, the results shown in Fig. 6a-b represent the global performance of a forensic system based on the formant trajectories of Spanish VS when considering all our speakers together, without separating them per speaker type (MZ, DZ, B) and following the leave-one-out cross-validated procedure described in Section 2.4.3. In the discussion section, we provide a brief comparison of these C_{llr} results with those obtained in similar studies in order to emphasize the comparatively good performance of our forensic system in a typical forensic scenario (i.e. without distinguishing MZ, DZ and B speakers). In the remaining part of this paper, we show only the results for the geometric mean, as the logistic regression fusion did not outperform the geometric mean fusion technique.

The Tippett plot in Fig. 7 represents the cumulative distribution of LLRs from two types of comparisons: same-speaker comparisons in the red line, and different-speaker comparisons in the blue line. In a first step, for different-speaker comparisons all the speakers have been pooled together, i.e. not distinguishing per speaker type. This represents the overall performance of our fused-19-VS system in a typical forensic scenario where only same-speaker and different-speaker comparisons are made. Some errors can be observed corresponding to contrary-to-fact LLRs. For instance, all different-speaker comparisons should obtain LLRs less than 0. However, the blue line invading the right quadrant beyond the green axis shows than some pairs of different speakers obtained positive LLRs. The opposite can be observed for the same-speaker comparisons: the red line which invades the left quadrant beyond the green axis represents some cases of comparing a speaker with himself (first session versus second session) and obtaining negative LLRs. Apart from these few cases, the Tippett plot shows a very good performance of our fused system when it comes to the standard forensic question "How much more likely are the observed differences between the known and questioned voice samples to occur under the hypothesis that the questioned sample has the same origin as the known sample than under the hypothesis that it has a different origin?" The incidence of contrary-to-fact support for

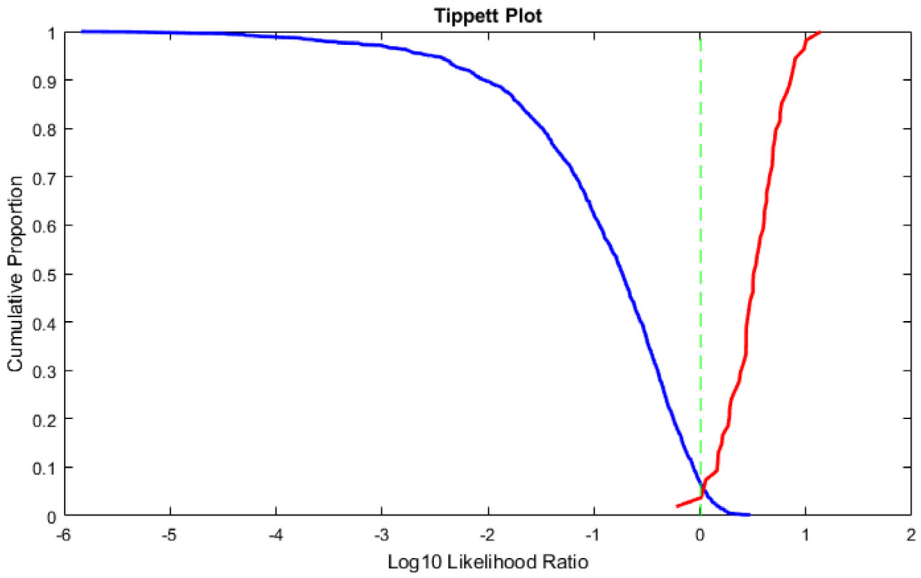

Fig. 7. Tippett plot for all pooled speakers. Red line = cumulative distribution of LLRs greater than or equal to the value indicated in the x-axis, calculated for same-speaker comparisons. Navy blue line = cumulative distribution of LLRs less than or equal to the value indicated in the x-axis, calculated for different-speaker comparisons (not distinguishing between MZ, DZ and B twins). Results for the DCT3 curve fitting technique and geometric mean fusion. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the same-speaker or different-speaker hypotheses is not very high, which corresponds with a good performing system, as the C_{llr} results revealed too.

Since one of our research questions was 'how do our different types of related speakers (MZ, DZ and B) affect the performance of a forensic-comparison system based on the VS of formant trajectories?', we present two Tippett plots (Fig. 8a-b), which show the LLR cumulative distribution according to the type of speaker comparison, for DCT3 and POLY3 correspondingly. Same-speaker comparisons are represented in the line rising to the right (red line). This is exactly the same as in Fig. 7, as intra-speaker comparisons do not change, but for the lines rising to the left we provide four different lines corresponding the four types of different-speaker comparisons: MZ, DZ, B and US comparisons.

On the one hand, Fig. 8a-b reveal again that the system performs quite well when considering a standard forensic scenario with unrelated speakers as non-targets and same speakers as targets. If we compare system performance tested on our three groups of related speaker pairs, we observe that contrary-to-fact LLRs occur mostly in the MZ group (black line). These are different-speaker comparisons for which the system yield a few LLRs that are larger than 0, pointing to the support of the same-speaker hypothesis. Comparatively, fewer contrary-to-fact cases can be found for the other two speaker groups: DZ (cyan) and B (magenta), which align with the distribution that we found for US (navy blue line). To a certain extent, this could imply that there is indeed higher intra-pair similarity in the formant dynamics of MZ twins than in other speaker comparisons (DZ, B or US). However, due to the nature of the graph, it is difficult to find support for the expected decreasing scale MZ > DZ > B > US. It is worth highlighting the uneven jagged aspect of all the non-target lines, as a clear indication of the small number of comparisons in each speaker category. This probably points to the inadequacy of this type of graph to answer fully to our first research question. For that reason, next section tackles this question in more detail and with different methodologies.

3.3. A comparison of MZ, DZ, B and US tests

One of our research questions revolved around whether there are statistically significant differences between the results extracted from comparing the formant dynamics of MZ twins and the results derived from other speaker comparisons (DZ, B or US). Considering that the LRs for intra-speaker (IS) comparisons should be higher than for MZ, the following decreasing scale in LRs would be expected: IS > MZ > DZ > B > US. Table 4 shows the values of two measures of central tendency (mean and median), together with the standard deviation, for the two types of curve fitting methods. Looking at the mean, we observe that the expected scaling in decreasing order from IS to US occurs always when considering both Poly3 and DCT3 except for the group of brothers, with higher values than DZ twins. This result is due to a single pair of brothers (see Appendix; Table A1). Instead of showing the results pooled for all the speakers of the same type, Table A1 in the Appendix shows all the individual comparisons and reveals that the high mean value of the B group in Table 4 is basically due to the pair of brothers 23–24, with strikingly high similarity. Discarding this outlier, the values for the rest of non-twin siblings are lower than the values obtained by DZ twins, as expected. In the discussion section, we provide some possible explanations for the unexpected values obtained by this pair of brothers.

Looking at the skewed distribution of values for the B group (Fig. 9), the median seems to be a better measure of central tendency for our data. Indeed, the LRs obtained for DZ pairs

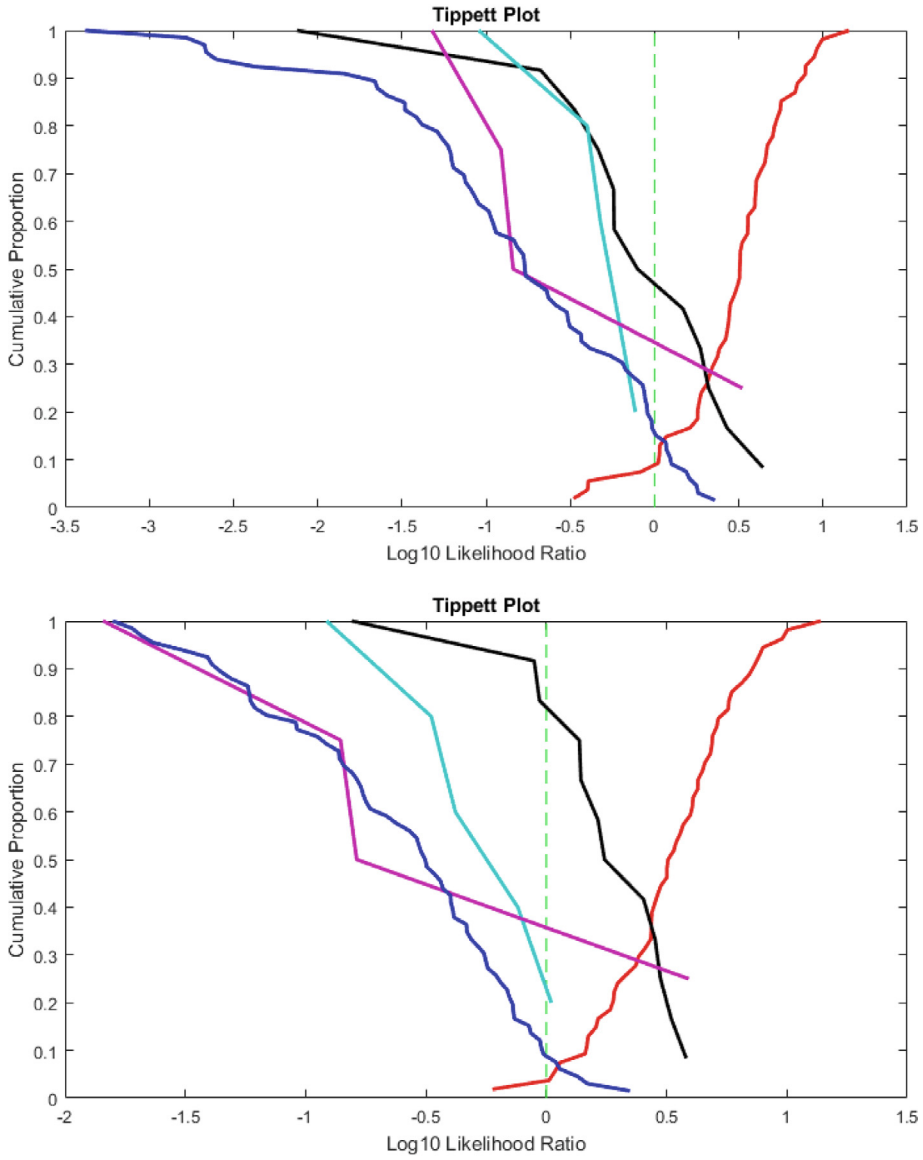

Fig. 8. a-b. Tippett plots showing the cumulative distribution of LLRs after geometric mean fusion. Red is used for same-speaker comparisons and navy blue for different-speaker (US) comparisons. The other three lines rising to the left represent one of the following IP comparisons: black is for MZs, cyan for DZs and magenta for B. Fig. 8a (top): results for the third-order polynomial curve fitting method; Fig. 8b (bottom): results for the DCT curve fitting method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
LR mean and standard deviation according to the type of curve fitting method and type of speaker comparison (IS = Intra-Speaker; MZ = monozygotic; DZ = dizygotic; B = brother; US = unrelated speakers), using the geometric-mean fusion method.

Type of comparison	Curve fitting	Mean	Median	Std. dev.
IS comparisons (*N* = 54)	Poly3	3.81	3.21	2.69
	DCT3	3.96	3.29	2.54
MZ comparisons (*N* = 12)	Poly 3	1.29	0.68	1.29
	DCT3	1.97	1.68	1.11
DZ comparisons (*N* = 5)	Poly3	**0.47**[a]	0.48	0.25
	DCT3	**0.54**	0.41	0.37
B comparisons (*N* = 4)	Poly3	0.91	*0.13*[b]	1.61
	DCT3	1.05	*0.14*	1.90
US comparisons (*N* = 66)	Poly3	0.44	0.17	0.54
	DCT3	0.42	0.30	0.42

[a] Results in bold show smaller values for DZ than for B comparisons.
[b] Results in italics show smaller values for B than for US comparisons.

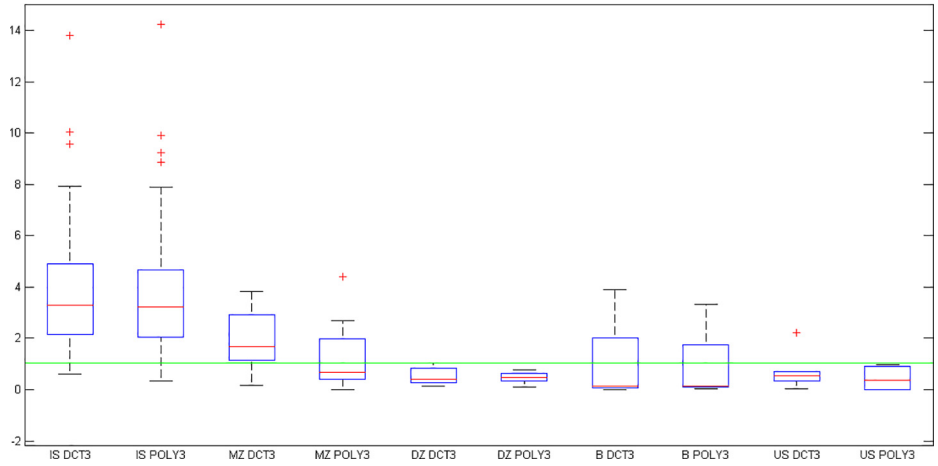

Fig. 9. Boxplots showing the distribution of LR values (combined under the geometric mean procedure) per type of comparison: IS (intra-speaker comparisons), MZ (monozygotic intra-pair comparisons), DZ (dyzigotic intra-pair comparisons), B (brother intra-pair comparisons) and US (unrelated-speaker intra-pair comparisons). The green line divides the graph in LRs > 1 and LRs < 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are higher than those obtained by B when the median is taken into account instead of the mean. This, however, results in the US group having higher LRs than the B group, against expectations (Table 4). The most plausible explanation for this is the different sample size (N) for each type of comparison. This N difference is particularly remarkable when comparing the group of brothers and the group of unrelated speakers.

The boxplots in Fig. 9 represent the distribution of the LRs shown in Table 4, according to the type of comparison (IS, MZ, DZ, B and US). Most values look normally distributed within each group with the notable exception of brothers, with an obvious skewed distribution. This is due, as explained before, to a specific pair (23–24), with values located at the whisker of the bloxplot, which would overlap not even with the average values of MZ comparisons but with the average values of IS comparisons, thus showing a striking similarity for this sibling pair.

Despite this, a decreasing scale can be observed in the expected direction from IS to US: IS > MZ > DZ > B > US. Even if considering the mean this trend was not observable, looking at the median (which is the value shown in boxplots), the expected trend appears more clearly. The green line in Fig. 9 divides the graph in LRs > 1 (support for the same-speaker hypothesis) and LRs < 1 (support for the different-speaker hypothesis). Interestingly, this green line leaves almost all cases of same-speaker comparisons (IS) above the green line and all cases of US different-speaker comparisons below the green line, with just an oulier, corresponding to speakers 27–28, with a LR = 2.22 for DCT3 (see Table A1). This is a further example of the good performance of the system when it is not challenged with related speakers. As explained in more detail in Section 3.2, our system does not give a high number of errors: contrary-to-fact hypothesis support is almost neglectable.

Apart from the IS and US LR-distribution at the two extremes of Fig. 9, we find the LR-distribution of MZ, DZ and B speaker pairs. Interestingly, for MZ the values obtained with the DCT3 curve fitting method are positive and those obtained with the POLY3 method are negative. However, they lie so close to 1 (i.e. the green line) that they do not provide a strong support for either H_{ss} or H_{ds}. This result supports the idea that this type of twins imply a great challenge for a forensic-comparison system. As for DZ and B, their mean values were found to be very similar in Table 4; so are their medians. Their values are not very different from the average US-comparison values, suggesting that, in general, they do not damage system performance as much as MZ speakers. See Section 4.3 for a discussion of these results and how they answer RQ-1b.

With the aim of finding whether there were statistically significant differences in terms of LRs between the five speaker-comparison types (IS, MZ, DZ, B and US), we used a Kruskal-Wallis H test (non-parametric test) after testing the residuals for normality. For the POLY3 data, the Kruskal-Wallis H test showed that there was a statistically significant difference in LRs between the different types of speaker comparisons, $\chi^2(4) = 82.37$, $p = 5.47 \times 10^{-17}$, with a mean rank LR of 70.75 for MZ, 53.20 for DZ, 49.75 for H, 42.27 for US and 109.39 for IS. For the DCT3 data, the Kruskal-Wallis H test showed that there was also a statistically significant difference in LRs between the different types of speaker comparisons, $\chi^2(4) = 94.98$, $p = 1{,}15 \times 10^{-19}$, with a mean rank LR of 86.58 for MZ, 49.60 for DZ, 42.38 for H, 39.31 for US and 110.37 for IS.

A series of pairwise Mann-Whitney U tests were conducted to test between which groups there were statistically significant differences. Regardless of the curve fitting method used, these tests showed that the LRs of IS comparisons are statistically significantly higher than the LRs of MZ, DZ and US (see Tables 5 and 6 for the specific statistical results). Interestingly, they are not significantly higher than the LRs of the B group (See discussion in Section 4.3). A further Mann-Whitney U test showed that the LRs of MZ comparisons –when using the DCT3 curve fitting method– are statistically significantly higher than the LRs of US ($U = 66$, $p = 5 \times 10^{-6}$; see Table 6).

Table 5
Mann-Whitney U test results (IS = Intra-Speaker; MZ = monozygotic; DZ = dizygotic; B = brother; US = unrelated speakers). Data for POLY3 curve fitting.

	IS	MZ	DZ	B	US
IS					
MZ	$U = 104$, $p = .0002^{*}$				
DZ	$U = 10$, $p = .0007^{*}$	n.s			
B	n.s	n.s	n.s		
US	$U = 132$, $p = 3.21 \times 10^{-18*}$	n.s	n.s	n.s	

n.s = non-significant results.
* = significant results (with Bonferroni correction).

Table 6
Mann-Whitney U test results (IS = Intra-Speaker; MZ = monozygotic; DZ = dizygotic; B = brother; US = unrelated speakers). Data for DCT3 curve fitting.

	IS	MZ	DZ	B	US
IS					
MZ	$U = 146$, $p = .003^{*}$				
DZ	$U = 3$, $p = 3.27 \times 10^{-4*}$	n.s.			
B	n.s.	n.s.	n.s.		
US	$U = 42$, $p = 4.45 \times 10^{-20*}$	$U = 66$, $p = 5 \times 10^{-6*}$	n.s.	n.s.	

n.s. = non-significant results.
* = significant results (with Bonferroni correction).

4. Discussion

4.1. Curve fitting

On the one hand, the results of the goodness-of-fit calculation showed that third-order functions fitted the trajectories better than the second-order functions. This occurred for all the VS and irrespective of the formant considered, implying that the adding of more coefficients for the curve fitting implies a more detailed or accurate approximation. This gives an answer to our third research objective (RO-3), which aimed to test whether a certain procedure for curve fitting outperforms the other.

On the other hand, we found a better goodness of fit in F2 than in F3. Although one can be tempted to attribute this to the fact that F2 is more constrained by the linguistic system while F3 is traditionally considered more speaker-specific (e.g. Battaner et al., 2003), this does not seem to be a completely plausible explanation, as each speaker's F contours are actually modelled separately. We think that this result may be due to the fact that F3 curves seem to have more inflection points and would therefore require higher order equations for a better fitting. In Appendix B we have included two examples of VS in which we can observe more inflection points in F3 than in F2 trajectories, although this is somehow VS-dependent, as we explain below. All in all, this result would imply that a few of the idiosyncratic F3 shapes fall outside the type of parametric curves explored for this study. Nevertheless, this does not affect our investigation greatly since both formants were combined into the MVKD formula for the joint characterization of each speaker. This was performed following previous studies (Morrison, 2009b) and with the aim of making a forensic-comparison system more powerful.

No clear trends were found as to whether some VS were better fitted than others. Although results varied depending on the formant and on the parametric function, correlation values were overall very high, indicating an accurate curve approximation. The most notable result was that /uo/ always obtained the lowest R values, as compared with the other VS (see Fig. A2 in the Appendix B). This happened irrespective of the type of parametric function or degree. Further studies would be necessary in order to explain the cause of these low correlation values. In contrast, we could not find a unique VS whose curve fitting was remarkably better than the others. Indeed, there were several VS with relatively high R values: /ai/, /ie/, /ue/ and /oe/ got values close to 1. The heterogeneity of this set of VS does not allow us to conclude that certain VS (e.g. rising diphthongs) are better correlated than others by means of the parametric functions used.

4.2. System performance

Three combination techniques were proposed to fuse the scores (of each of the 19 systems, one per VS) obtained after applying the MVKD formula: 1) naïve Bayes; 2) geometric mean; and 3) logistic regression. The cost function C_{llr} allowed us to evaluate the performance of our fused systems as well as that of the individual systems (considering the VS separately). The results showed that the use of fusion techniques (either geometric mean or logistic-regression fusion) improved system performance: any fused system outperforms any of the systems based on individual VS. The best-performing system is therefore the one based on geometric-mean sum and DCT3 curve fitting (0.15 C_{llr}). The implication is that the best-performing forensic system is one that characterizes each speaker by all their VS, not individual ones. This responds to our second research objective (RO-2), aimed at testing for the first time whether the fusion of the 19 VS of Spanish outperforms the individual VS for speaker comparison.

Our results are in line with those obtained by González-Rodríguez et al. (2007) and Gil-Gil (2009) for other languages. They also compared both types of combination procedures

and did not find that the method which implies calibration (logistic regression) provides better results than the technique which assumes statistical independence (like the geometric mean procedure). Indeed, González-Rodríguez et al. (2007) and Gil-Gil (2009) combine the scores from systems based on different diphthongs, as in our study. In contrast, in their study of the formant trajectories of the monophthong /o/, Morrison and Kinoshita (2008) concluded that substantial improvement was found in system performance when the output of MVKD was calibrated using logistic regression. The implication of our findings is, in agreement with González-Rodríguez et al. (2007), that we can find a reasonable independence of the contribution of the different diphthongs to the voice evidence.

In case of time constraints preventing the extraction of all the VS of a speaker in forensic casework, it seems that certain vowel combinations give better results than others. Regardless of the curve-fitting method, we found the following trends. On the one hand, VS involving two back vowels (/uo/) or the combination of mid-back (/ao/) and back-mid (/ua/ and /oa/) vowels give a consistently lower C_{llr} than the VS with the combination front + back and close vowel (/eu/, /iu/), which seem to perform slightly better. Presumably, this is due to the minimal dynamic movement in /uo/, /ua/ and /ao/ while in /eu/ and /iu/ there is an opportunity for more individuality in how speakers move between the targets across the vowel space. Compare these results with the good performance reported by McDougall (2006) in Australian English /aɪ/. There are also differences depending on the curve fitting method: /ea/ is the best-performing VS for POLY3 and /oi/ for DTC3. Since /ao/ and /oa/ tend to have a hiatus realization (see Table 1) while /eu/ and /iu/ are typically realized as diphthongs, in future studies we will consider the separate study of diphthong-hiatus realizations for all VS. Surprisingly, /ua/ appears consistently –across curve fitting methods– as the worst-performing VS. Some extra-words containing this VS were included, expecting high between-speaker variation and better speaker-identification potential. However, it seems that such expected pronunciation oscillations affect also some speakers internally. In other words, alongside inter-speaker variation we also observe high intra-speaker variation in the realization of these sequences when comparing non-contemporaneous speaking sessions.

All in all, the post-calibration C_{llr} values obtained in our study (ranging between 0.15 and 0.30 depending on the curve fitting method) are comparable to or even better than those obtained in previous studies of the same nature. The best formant-based fused system in Franco-Pedroso & Gonzalez-Rodriguez (2016) obtains a C_{llr} of 0.374 for male speakers (English-only trials) while Morrison (2009b) provides a C_{llr} of 0.218 for a fused system comprising five Australian English diphthongs (MKVD results). This responds to our first research objective (RO-1), aimed at testing the forensic-phonetic potential of formant trajectories extracted from Spanish VS. In particular, the first research question (RQ-1a) was: what is the overall performance of a forensic-comparison system based on VS formant trajectories? The answer would be that the best-performing system is one based on the geometric-mean sum of scores and DCT3 curve fitting of 19 VS (0.15 C_{llr}).

The Tippet plots in Fig. 8 a-b serve to answer the second research question (RQ-1b): how do our different types of related speakers (MZ, DZ and B) affect the performance of the system? The results of our investigation show a deterioration in performance particularly when we compare MZ twin pairs. DZ and B twin pairs do not seem to affect performance so clearly, as LLRs align with the values obtained by other different-speaker comparisons (i.e. those comparing unrelated speakers). Due to the small size of the MZ, DZ and B subsets, a C_{llr} was not provided but the results provided in Fig. 8 a-b and Table A1 allow us to conclude that performance testing with twins is very useful to highlight the system errors when very similar-sounding speaker pairs are compared.

4.3. A comparison of MZ, DZ, B and US tests

Once we tested that MZ, DZ and B comparisons clearly affect system performance, we wondered whether there would be statistically significant differences between the results obtained by MZ comparisons and those obtained after comparing other related speakers. Due to the different sample size of the different speaker groups considered, neither the mean values nor the median values allowed us to observe the expected decreasing scale in LRs completely clearly: IS > MZ > DZ > B > US, although this trend was clearly observable in the boxplots, regardless of the type of function considered (Poly3 or DCT3). Precisely due to the small size of the brother group (four non-twin sibling pairs participated) it was possible to detect in Table A1 the origin of the discordant mean value of this group, preventing the expected decreasing scale. The comparison of brother pair 23–24 yielded relatively high LRs: 3.32 (Poly3) and 3.90 (DCT3). These values are very close to those of the most similar MZ pairs, and even close to typical intra-speaker comparison values. The fact that only a pair of non-twin siblings shows such high values, in comparison with the rest of brothers, makes the standard deviation of this group very large. This striking value makes the distribution of the brother group very skewed, as can be clearly seen in the boxplots.

In general, it has to be said that LRs around 1 or LLRs around 0 do not indicate a strong support for either of the competing hypotheses. This is the case of MZ comparisons, so it would be difficult for the forensic-comparison system to decide whether the two speech samples come from the same or from different speakers. The fact that the system cannot tip the balance in favor of one hypothesis or the other when comparing MZ twins is in agreement with the fact that these speakers are very similar. Depending on the specific twin pair considered, higher or lower LRs are yielded by the system. This would indicate that the parameters considered are not uniquely and completely genetically influenced. On the contrary, non-genetic aspects (like learned habits) should be exerting a strong influence. In other words, the fact that some MZ twin pairs get high LRs while other get lower values suggest that factors like the ones considered in the questionnaire should be taken into account to explain the variation. We refer to factors such as degree of relationship closeness, shared /non-shared leisure activities, shared/non-shared group of friends, time spent together, and so on.

When considering the B group in particular, we have already marked how heterogeneous this group is. The high LRs of one specific pair (23–24) suggest that the phonetic parameters studied are, to a high degree, environmentally influenced. With this we mean that the phonetic realization of the VS formant trajectories must be strongly subject to the specific and voluntary implementation of the acoustic target by the speaker, naturally within his anatomical constraints. The different behavior of the B pair 23–24 in comparison with the rest of speakers in his group can be explained by several factors. We have looked in detail at the responses given by these speakers in the questionnaire gathered at the first speaking session, and the following remarkable aspects are to be noted:

- Speakers 23 and 24, with an age difference of 7 years, are the only ones among the B speakers who answer "Very often" to the question "How often do people confuse your voice with that of your brother?"
- They are the only ones in his group answering "Absolutely not. I think that we speak the same way" to the question "Do you consider that your voice/manner of speaking is very different from your brother's?"
- They are the only ones among the B speakers who share leisure activities. In comparison with the rest of non-twin brothers, speaker 23 and 24 see each other quite often (at least once a week) and they also talk to each other quite often (between twice and three times per week).
- In the open question "Mention a few aspects in which you think (or people have commented about how) your voice is different/similar from your brother's", they answer that many people have mentioned their same way of laughing, their similar intonation and their use of similar expressions. Besides, one of them mentions the anecdote of having been once recognized as brothers by a certain common acquaintance solely on the basis of their voice, without being both of them ever together before that person and without this person having beforehand knowledge of their family kinship.
- Finally, in the question related to their degree of closeness, from 1 to 5 (being 1 "not very close" and 5 "very close"), they gave 4.5 points on average.

All of the above-mentioned responses given in their questionnaires could be indicative of the nurture factors outweighing the genetic ones for explaining the strikingly high LRs values obtained in their comparison. Furthermore, in a perceptual study in which these same speakers participated (San Segundo, 2013b), the laughter of these brothers was actually found to be very similar to each other. The question of whether this was due to a similar vocal tract or to imitated behavior was not tackled. Another plausible explanation –rather than nurture outweighing nature– could be that these two factors have their own bearing on the high similarity of this sibling pair. In relation to genetic aspects (i.e. nature), it should be borne in mind that siblings –just the same as DZ twins– share approximately 50% of their genes in common but in the case of same-sex pairs a realistic range is probably from 25% to 75% of the total genome (Pakstis et al., 1972). As these authors state, a pair with more genes in common should be more similar in appearance and behavior than those with fewer genes in common. In any case, in view of the results of their questionnaires, we can conclude that this is a case of a non-twin sibling pair having a closer relationship than many of the MZ or DZ twins also participating in this study, which would have clearly exerted certain "intra-sibling mimetism" in the speech patterns of these brothers.

All in all, there seems to be two opposite directions which the relationship between siblings can head for: towards accommodating or towards distancing in their speech behavior. In this study, the accommodating effect may have been reinforced by the type of speaking task from which the phonetic parameters were extracted. As it was an information exchange between conversational partners (i.e. a collaborative exercise), this may have triggered certain convergence in the speech habits of the speakers, possibly resulting in similar acoustic outputs for the VS considered. In this respect, our investigation differs from previous studies on twins and formant dynamics, which are either based on wordlist reading tasks (Zuo & Mok, 2015) or on Labovian-style interviews with the researcher (Loakes, 2006). To the best of our knowledge, this is the first phonetic investigation on twins which extracted formant dynamic information from informal exchanges between twin pairs and non-twin sibling pairs. As the Twin Corpus also includes semi-structured interviews between the researcher and each sibling separately, in future studies it would be interesting to examine the extent to which accommodation effects are observed in different types of speaking tasks. For further discussions about convergence and imitation patterns in speech occurring between speakers in the course of conversational interactions, see Babel (2009), Coupland (1984), Giles, Coupland and Coupland (1991), Pardo (2006), Pickering and Garrod (2004) or Pardo et al. (2012).

If we focus on twins alone, it has to be noted that all our twin participants (MZ and DZ) were reared together. However, previous studies have revealed that twins reared together can sometimes be more different than twins reared apart. Although this may sound surprising, it has been suggested that "twins reared together may 'create' differences between themselves in an attempt at differentiation from the twin" (Segal, 1990: 615). As regards the question of whether MZ twins have a closer relationship than DZ twins, it seems that "an impressive body of experimental, clinical, and observational data suggests that MZ twins share a more intimate social bond, relative to DZ twins" (Burlingham, 1952; Mowrer, 1954; Paluszny et al., 1977; Segal, 1984; Smith, Renshaw & Renshaw, 1968; In Segal, 1990: 619). This fact could be at the base of what Debruyne, Decoster, Van Gysel, and Vercammen (2002) call "intratwin mimetism".

From the results of the pairwise Mann-Whitney U tests, two main conclusions can be drawn. Firstly, the fact that there are no significant differences between the LRs obtained in the comparison of MZ, DZ, B and US pairs points to the good performance of the forensic system proposed here (consisting of all the 19 VS of Spanish). Even though we saw in the Tippett plots that all related speaker pairs (MZ, DZ and B) affect somehow system performance, the LRs obtained by MZ, DZ and B are not significantly higher than those of US, which would be the kind of different speakers tested in a typical forensic scenario. This implies that MZ, DZ and B comparisons on average give factual support to the different-speaker hypothesis. Note that with a DCT3 approach, the MZ group is the only one

among related speakers with LRs which are significantly higher than US. This supports what we said before about the challenge that MZ tests imply for forensic systems, although further investigations would be necessary to explain why this happens only with the DCT3 approach and not with the POLY3. Secondly, regardless of the curve fitting method, these Mann-Whitney U tests showed that the LRs of IS comparisons are statistically significantly higher than the LRs of MZ, DZ and US, as it is expected and desirable for a good-performing system. The fact that B tests are the only different-speaker group where this result is not obtained –regardless of the fact that this result is due to a strikingly similar brother pair– further supports the importance of challenging forensic systems with related speakers of different degrees, apart from MZ twins.

5. Conclusions

We have investigated formant dynamics in Spanish VS and shown that they have great potential for FVC. This is most probably due to the genuine distinction of Spanish between hiatuses and diphthongs, and to the fact that a number of pronunciation vacillations are allowed within the linguistic system, favoring idiosyncratic pronunciations. We have undertaken this investigation from a two-fold perspective. On the one hand, we have investigated the overall performance of a forensic-comparison system consisting of fused and unfused VS. On the other hand, we have explored the question of how related speakers (MZ and DZ twins, as well as brothers) affect the performance of our system.

Firstly, the results of our study show that the fusion of 19 VS outperforms the individual VS for speaker comparison, and that the geometric-mean combination method outperforms logistic regression – when using the Multivariate Kernel Density Formula to fuse F2 and F3 dynamic coefficients. Finally, as expected, third-degree Discrete Cosine Transform (DCT3) and cubic polynomial functions outperform their second-degree counterparts as curve fitting methods for the examined formant trajectories. The overall performance of an all-VS system is very good in a classic forensic scenario which does not consider related speakers (C_{llr} = 0.15 with the DCT3 curve fitting method and the geometric-mean fusion method).

Secondly, as far as the question of related speakers is concerned, MZ twin pairs deteriorate the performance of the all-VS comparison system that we have tested. More importantly, however, our results also show that brother comparisons can affect system performance too, and to a great extent. This suggests that more investigations are necessary into a range of related, similar-sounding speakers to challenge forensic comparison systems beyond MZ twins. Nurture factors were the most plausible explanation for the strikingly high similarity found in a specific non-twin sibling pair, as shown by their questionnaire responses. These results are in line with the explanation that previous authors (Loakes, 2006 and Zuo & Mok, 2015) have found for the (dis)similarities of several twin pairs. As in the case of twins, learned variation, individual choice and the attitude towards one's own sibling seem to play an important role in speech production and can explain the extremely convergent formant patterns in a non-twin sibling pair.

All in all, these results point to the importance of undertaking more phonetic studies into the speech similarities of brothers and not only twins, particularly in forensic and biometric applications. Twin investigations could be criticized as being too exotic, probably due to the low incidence of MZ twins worldwide (approximately four MZ twin births per thousand births). For that reason twins are difficult to recruit and scarce twin voice databases exist, not to mention twin registries, available only in some countries. Without neglecting the importance of undertaking twin studies –it is indeed important to focus on special populations for answering particular questions– the results of this investigation point to the adequacy of testing the performance of forensic systems in other types of related speaker pairs too. Phonetic investigations into the speech of non-twin siblings are rare (Kinga, 2007), particularly if undertaken within a forensic (see Section 1.1) or a biometric perspective (Charlet & Peral, 2007). This is particularly surprising given that the incidence of brothers in the population is clearly higher than that of twins –making them easier to recruit– and given the fact that it is not uncommon to find forensic cases where relatives, particularly brothers, are involved. From a sociophonetic perspective, the study of related speakers is also important in order to understand to what extent the family, as one of the most basic units in society (Benson & Deal, 1995; Hazen, 2002), affects the phonetic output of the individual and exerts important effects on language variation patterns in the same or in a different way that other peer group interactions.

Acknowledgements

This work was partly supported by the Spanish Ministry of Science and Innovation: *Programa de Formación de Profesorado Universitario* [grant number AP-2008-01524] and partly thanks to the financial support of the innovation team of the Shanxi Police College. Eugenia San Segundo thanks Prof. Joaquim Llisterri and Dr. Daniel Ramos for insightful feedback to earlier versions of this investigation, as part of her doctoral dissertation. We also thank two anonymous reviewers for their useful comments on this paper.

Appendix A

Table A1

Table A1
LRs obtained per type of comparison (IS = intra-speaker; IP = intra-pair) and per curve fitting method (Poly3 = cubic polynomial function; DCT3 = third-degree DCT function); xxvyy means speaker xx versus speaker yy. The fusion method is geometric mean.

	Monozygotic twins			Dizygotic twins			Brothers			Unrelated speakers		
	IS-test		IP-test	IS-test		IP-test	IS-test		IP-test	IS-test		IP-test
	01v01	02v02	01v02	13v13	14v14	13v14	21v21	22v22	21v22	25v25	26v26	25v26
Poly3	*5.33*[a]	*3.95*	**4.41**[b]	*8.87*	*2.05*	**0.61**	*9.91*	*7.83*	**0.14**	*4.67*	*2.78*	**0.54**
DCT3	*4.27*	*3.47*	**2.83**	*4.05*	*2.73*	**1.05**	*13.82*	*7.94*	**0.14**	*4.07*	*2.58*	**0.34**
	03v03	04v04	03v04	15v15	16v16	15v16	23v23	24v24	23v24[c]	27v27	28v28	27v28
Poly3	*3.22*	*2.82*	**0.58**	*2.64*	*5.57*	**0.77**	*4.60*	*3.23*	**3.32**	*3.59*	*5.03*	**0.91**
DCT3	*5.19*	*2.90*	**1.75**	*1.98*	*4.91*	**0.33**	*5.91*	*6.50*	**3.90**	*3.39*	*9.57*	**2.22**
	05v05	06v06	05v06	17v17	18v18	17v18	47v47	48v48	47v48	29v29	30v30	29v30
Poly3	*2.12*	*1.17*	**0.46**	*1.84*	*0.40*	**0.48**	*1.89*	*4.00*	**0.05**	*2.75*	*3.58*	**0.00**
DCT3	*1.64*	*1.90*	**0.94**	*3.17*	*1.92*	**0.42**	*1.48*	*3.17*	**0.01**	*1.08*	*0.60*	**0.40**
	07v07	08v08	07v08	19v19	20v20	19v20	49v49	50v50	49v50	31v31	32v32	31v32
Poly3	*7.90*	*5.13*	**0.34**	*6.84*	*9.24*	**0.40**	*2.97*	*1.62*	**0.12**	*3.94*	*3.20*	**0.01**
DCT3	*5.70*	*3.21*	**0.16**	*5.10*	*7.65*	**0.76**	*2.73*	*1.49*	**0.16**	*6.99*	*2.75*	**0.02**
	09v09	10v10	09v10	45v45	46v46	45v46				51v51	52v52	51v52
Poly3	*1.07*	*0.82*	**0.79**	*0.33*	*3.10*	**0.09**				*3.31*	*3.19*	**0.96**
DCT3	*1.14*	*1.03*	**0.89**	*1.45*	*4.81*	**0.12**				*1.84*	*7.33*	**0.70**
	11v11	12v12	11v12							53v53	54v54	53v54
Poly3	*1.80*	*1.07*	**0.21**							*3.57*	*4.24*	**0.65**
DCT3	*2.99*	*2.36*	**1.37**							*1.61*	*2.82*	**0.17**
	33v33	34v34	33v34									
Poly3	*4.52*	*5.64*	**1.87**									
DCT3	*3.96*	*4.88*	**3.82**									
	35v35	36v36	35v36									
Poly3	*1.05*	*2.36*	**2.69**									
DCT3	*3.71*	*2.72*	**2.98**									
	37v37	38v38	37v38									
Poly3	*0.40*	*4.01*	**2.09**									
DCT3	*4.57*	*3.61*	**3.30**									
	39v39	40v40	39v40									
Poly3	*2.43*	*14.25*	**0.01**									
DCT3	*2.16*	*5.74*	**1.64**									
	41v41	42v42	41v42									
Poly3	*2.70*	*2.22*	**1.47**									
DCT3	*4.26*	*2.42*	**1.39**									
	43v43	44v44	43v44									
Poly3	*1.80*	*7.04*	**0.57**									
DCT3	*4.45*	*10.03*	**2.54**									

[a] The LRs of all IS-tests are highlighted in italics. These are intra-speaker comparisons (first session vs. second session of each speaker).
[b] The LRs of all IP-tests are highlighted in bold. These are intra-pair comparisons between monozygotic twins, dizygotic twins, brothers and unrelated speakers.
[c] Pair of brothers 23v24, with LRs strikingly high for a non-twin sibling pair (compare with the results obtained in the monozygotic IP-tests). See discussion, Section 4.3.

Appendix B

Figs. A1 and A2

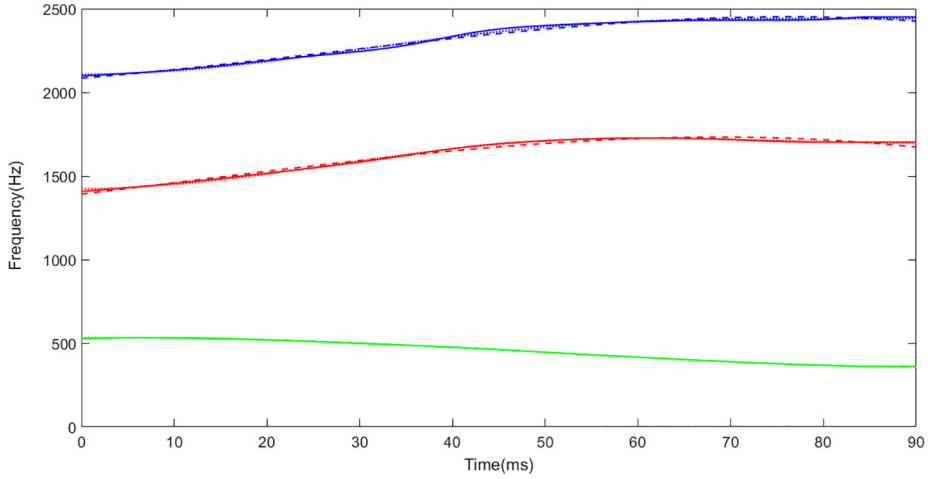

Fig. A1. Example of curve fitting: approximation of the formant trajectories (F1 to F3) of the VS /ai/ extracted from the word *países* pronounced by speaker 49. Green: F1; red: F2; blue: F3. Continuous line: original formant trajectories; dashed line: 3rd-degree polynomial approximation; dotted line: 3rd-degree DCT approximation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

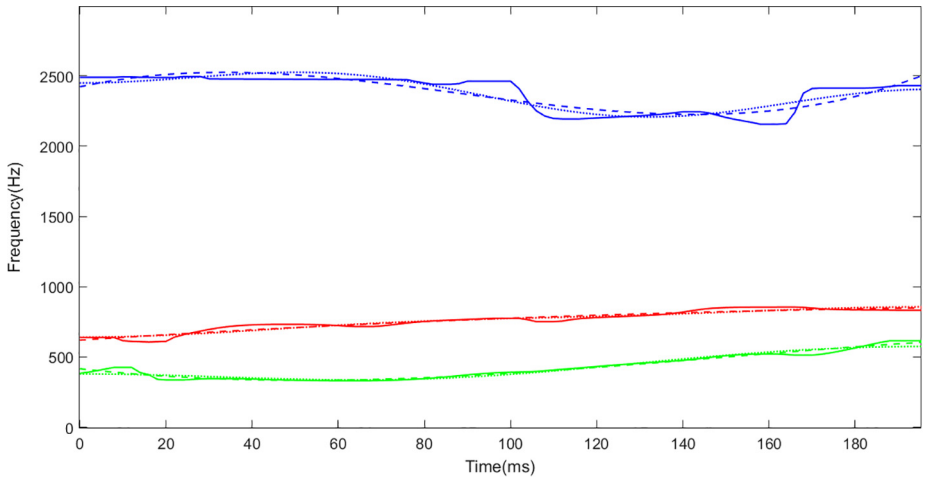

Fig. A2. Example of curve fitting: approximation of the formant trajectories (F1 to F3) of the VS /uo/ extracted from the word *búho* pronounced by speaker 23. Green: F1; red: F2; blue: F3. Continuous line: original formant trajectories; dashed line: 3rd-degree polynomial approximation; dotted line: 3rd-degree DCT approximation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

Abril, A., Ambrosio, E., de Blas, M., Caminero, A., García, C., & de Pablo, J. (2009). *Fundamentos de psicobiología*. Madrid: Sanz y Torres.

Aguilar, L. (1999). Hiatus and diphthong: Acoustic cues and speech situation differences. *Speech Communication, 28*(1), 57–74.

Aguilar, L. (2010). *Vocales en grupo*. Madrid: Arco/Libros.

Aitken, C., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 53*(1), 109–122.

Alarcos Llorach, E. (1965). *Fonología española* (4th ed.). Madrid: Editorial Gredos.

Alves, H., Rico, J., & Roca, I. (2010). BuFón: Buscador de patrones fonológicos. Retrieved from: http://www.estudiosfonicos.cchs.csic.es/fonetica/bufon?p=presentacion. (Last accessed July 2014).

Anderson, S. (1985). *Phonology in the twentieth century*. Chicago: University of Chicago Press.

Anderson, A. H., Badger, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351–366.

Babel, M. (2009). *Phonetic and social selectivity in speech accommodation (Doctoral dissertation)*. Berkeley: Department of Linguistics, University of California.

Battaner, E., Gil, J., Marrero, V., Llisterri, J., Carbó, C., & Machuca, M., . . . & Ríos, A. (2003). VILE: Estudio acústico de la variación inter e intralocutor en español. In SEAF 2003: Actas del II Congreso de la Sociedad Española de Acústica Forense (pp. 59-70).

Benson, M., & Deal, J. (1995). Bridging the individual and the family. *Journal of Marriage and The Family*, 561–566.

Berger, C., Robertson, B., & Vignaux, G. (2010). Interpreting scientific evidence. In I. Freckelton & H. Selby (Eds.), *Expert Evidence*. Sydney: Thomson Reuters.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer [Computer software] (Version 6.0.42). Retrieved from http://www.praat.org.

Borzone de Manrique, A. M. (1979). Acoustic analysis of the Spanish diphthongs. *Phonetica, 36*(3), 194–206.

Bruder, C. E., Piotrowski, A., Gijsbers, A. A., Andersson, R., Erickson, S., de Ståhl, T. D., ... Crowley, M. (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *The American Journal of Human Genetics, 82*(3), 763–771.

Brümmer, N. (2005). FoCal: Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers. Retrieved from: https://sites.google.com/site/nikobrummer/focal. [Computer software].

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language, 20*(2), 230–275.

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., ... Strasheim, A. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(7), 2072–2084.

Bulmer, M. G. (1970). *The biology of twinning in man*. Oxford: Oxford University Press.

Burlingham, D. (1952). *Twins; a study of three pairs of identical twins*. New York: International Universities Press.

Cabré, T., & Prieto, P. (2006). Exceptional hiatuses in Spanish. *Optimality-theoretic studies in Spanish phonology, 99*, 205–238.

Calderwood, I. (2015, September 7). Mystery of which identical twin committed a series of rapes in France is finally solved as one brother confesses after he was given away by a stutter, Mailonline. Retrieved from https://www.dailymail.co.uk/news/article-3225467/Mystery-identical-twin-committed-series-rapes-France-finally-solved-one-brother-confesses-given-away-stutter.html.

Cambier-Langeveld, T. (2007). Current methods in forensic speaker identification: Results of a collaborative exercise. *International Journal of Speech, Language and the Law, 14*(2), 223–243.

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication, 31*(2–3), 193–203.

Charlet, D., & Peral, V. (2007). Voice Biometrics within the Family: Trust, Privacy and Personalisation. E-business and Telecommunication Networks (Second International Conference, ICETE 2005, Reading, UK, October 3-7, 2005. Selected Papers), 3, 93-100.

Colina, S. (1999). Reexamining Spanish glides: Analogically conditioned variation in vocoid sequences in Spanish dialects. In J. Gutiérrez-Rexach & F. Martínez-Gil (Eds.), *Advances in Hispanic Linguistics* (pp. 121–134). Somerville MA: Cascadilla.

da Costa Fernandes, V. S. (2018). *Alterações acústicas e percetivas introduzidas nas vozes de indivíduos gémeos e devidas ao canal telefónico - Uma discussão de impacto na análise forense (Doctoral dissertation)*. Universidad de Porto.

Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language, 46*, 49–70.

Debruyne, F., Decoster, W., Van Gijsel, A., & Vercammen, J. (2002). Speaking fundamental frequency in monozygotic and dizygotic twins. *Journal of Voice, 16*(4), 466–471.

Decoster, W., Van Gysel, A., Vercammen, J., & Debruyne, F. (2000). Voice similarity in identical twins. *Acta Oto-Rhino-Laryngologica Belgica, 55*(1), 49–55.

Enzinger, E. (2010). Characterising Formant Tracks in Viennese Diphthongs for Forensic Speaker Comparison. In *Proceedings of the 39th International AES Conference: Audio Forensics, Practices and Challenges* (pp. 47–52).

Evett, I., & Buckleton, J. (1996). Statistical analysis of STR data. In A. Carraredo, B. Brinkmann, & W. Bär (Eds.), *Advances in Forensic Haemogenetics* (pp. 79–86). Heidelberg: Springer-Verlag.

Feiser, H. (2009). Acoustic similarities and differences in the voices of same-sex siblings. *Paper presented at the 18th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA) for Forensic Phonetics and Acoustics (IAFPA)*. Cambridge, UK.

Feiser, H., & Kleber, F. (2012). Voice similarity among brothers: evidence from a perception experiment. *Paper presented at the 21st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*. Santander, Spain.

Forrai, G., & Gordos, G. (1982). A new acoustic method for the discrimination of monozygotic and dizygotic twins. *Acta Paediatrica Hungarica, 24*(4), 315–322.

Franco-Pedroso, J., & Gonzalez-Rodriguez, J. (2016). Linguistically-constrained formant-based i-vectors for automatic speaker recognition. *Speech Communication, 76*, 61–81.

Gedda, L., Fiori-Ratti, L., & Bruno, G. (1960). La voix chez les jumeaux monozygotiques. *Folia Phoniatrica et Logopaedica, 12*(2), 81–94.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in Applied Sociolinguistics*. Cambridge: Cambridge University Press.

Gil-Gil, J. (2009). Identificación forense de locutor mediante el empleo de relaciones de verosimilitud sobre secuencias vocálicas como función discriminante y uso de la entropía cruzada empírica como medida. *de precisión de los resultados (Master's thesis)*. Universidad Autónoma de Madrid.

Gold, E., & French, P. (2011). An international investigation of forensic speaker comparison practices. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 1254–1257). Hong Kong, China.

Goldstein, U. (1976). Speaker-identifying features based on formant tracks. *Journal of the Acoustical Society of America, 59*(1), 176–182.

González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D., & Ortega-García, J. (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing, 15*(7), 2104–2115.

Hall, J. G. (2003). Twinning. *The Lancet, 362*(9385), 735–743.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer.

Hazen, K. (2002). The family. In J. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (pp. 500–525). Malden, MA: Blackwell.
Himmelreich, C. (2009). Despite DNA Evidence, Twins Charged in Heist Go Free, Time. Retrieved from http://content.time.com/time/world/article/0,8599,1887111,00.html.
Hualde, J. I. (1991). On Spanish syllabification. In In H. Campos & F. Martínez Gil (Eds.), *Current Studies in Spanish Linguistics* (pp. 475–493). Washington: Georgetown University Press.
Hualde, J. I. (2005). *The sounds of Spanish*. Cambridge: Cambridge University Press.
Hualde, J. I., & Chitoran, I. (2003). Explaining the distribution of hiatus in Spanish and Romanian. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 3013–3016).
Hualde, J. I., & Prieto, M. (2002). On the diphthong/hiatus contrast in Spanish: Some experimental results. *Linguistics, 40*(2; ISSU 378), 217–234.
Jessen, M. (2008). Forensic phonetics. *Language and Linguistics. Compass, 2*(4), 671–711.
Johnson, K., & Azara, M. (2000). *The perception of personal identity in speech: Evidence from the perception of twins' speech*. Unpublished Manuscript.
Kinga, P. (2007). Hereditary phonetic parameters of the human voice. *Magyar Nyelvor (Hungarian Language Guardian), 131*(3), 306–315.
Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication, 52*(1), 12–40.
Kinoshita, Y., & Osanai, T. (2006). Within speaker variation in diphthongal dynamics: What can we compare? In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 112–117).
Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmissions on the measurement of formant frequencies. *Forensic Linguistics, 8*(1), 80–99.
Künzel, H. J. (2011). Automatic speaker recognition of identical twins. *International Journal of Speech, Language and the Law, 17*(2), 251–277.
Loakes, D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins (Doctoral dissertation)*. University of Melbourne.
Markel, J., & Gray, A. (1976). *Linear prediction of speech*. Berlin: Springer-Verlag.
Martínez-Paricio, V. (2013). The intricate connection between diphthongs and stress in Spanish. *Nordlyd, 40*(1), 166–195.
Matheny, A., & Bruggemann, C. (1973). Children's speech: heredity components and sex differences. *Folia Phoniatrica, 25*(6), 442–449.
McDougall, K. (2004). Speaker-specific formant dynamics: an experiment on Australian English /ai/. *International Journal of Speech Language and the Law, 11*(1), 103–130.
McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies. *International Journal of Speech Language and the Law, 13*(1), 89–126.
Morrison, G. S. (2007). Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. Available from http://geoff-morrison.net/#MVKD.
Morrison, G. S. (2009a). Forensic voice comparison and the paradigm shift. *Science & Justice, 49*(4), 298–308.
Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America, 125*(4), 2387–2397.
Morrison, G. S. (2010a). Forensic Voice Comparison. In I. Freckelton & H. Selby (Eds.), *Expert Evidence*. Sydney: Thomson Reuters.
Morrison, G. S. (2010b). Sound file cutter upper. [Computer software] Retrieved from: http://geoff-morrison.net/#CutUp.
Morrison, G. S. (2012). SoundLabeller: Ergonomically designed software for marking and labelling sections of sound files. [Computer software] Retrieved from: http://geoff-morrison.net/#SndLbl.
Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences, 45*(2), 173–197.
Morrison, G. S., & Kinoshita, Y. (2008). Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. In *Proceedings of Interspeech* (pp. 1501–1504).
Morrison, G.S., & Nearey, T. (2011). FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories. [Computer software] Retrieved from: http://geoff-morrison.net/#FrmMes.
Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences, 44*(2), 155–167.
Mowrer, E. (1954). Some factors in the affectional adjustment of twins. *American Sociological Review, 19*(4), 468–471.
Navarro Tomás, T. (1946). *Estudios de fonología española*. Syracuse, N.Y.: Syracuse University Press.
Navarro Tomás, T. (1918). Manual de pronunciación española. Madrid: Consejo Superior de Investigaciones Científicas, 1972 (17th ed.).
Nearey, T., Assmann, P., & Hillenbrand, J. (2002). Evaluation of a strategy for automatic formant tracking. *Journal of the Acoustical Society of America, 112*(5), 2323. 2323.
Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.
Nolan, F. (2002). The "telephone effect" on formants: a response. *Forensic Linguistics, 9*(1), 74–82.
Nolan, F., & Oh, T. (1996). Identical twins, different voices. *International Journal of Speech Language and the Law, 3*(1), 39–49.
Pakstis, A., Scarr-Salapatek, S., Elston, R. C., & Siervogel, R. (1972). Genetic contributions to morphological and behavioral similarities among sibs and dizygotic twins: Linkages and allelic differences. *Social Biology, 19*(2), 185–192.
Paluszny, M., Selzer, M., Vinokur, A., & Lewandowski, L. (1977). Twin relationships and depression. *The American Journal of Psychiatry, 134*, 988–990.
Pardo, J. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America, 119*(4), 2382–2393.
Pardo, J., Gibbons, R., Suppes, A., & Krauss, R. (2012). Phonetic convergence in college roommates. *Journal of Phonetics, 40*(1), 190–197.
Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–190.
Pigeon, S., Druyts, P., & Verlinde, P. (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing, 10*(1), 237–248.
Quilis, A. (1981). *Fonética acústica de la lengua española*. Madrid: Gredos.
Real Academia Española y Asociación de Academias de la Lengua Española (RAE). (2011). Nueva gramática de la lengua española. Fonética y Fonología. (pp. 332–354). Madrid: Espasa.
Ramos-Castro, D. (2007). *Forensic evaluation of the evidence using automatic speaker recognition systems* (Doctoral dissertation). Universidad Autónoma de Madrid.
Rose, P. (2002). *Forensic speaker identification*. London: Taylor & Francis.
Rose, P. (2003). The technical comparison of forensic voice samples. In Selby Freckelton (Ed.), *Expert Evidence 99* (pp. 1051–6102). Sydney: Thomson Reuters.
Rose, P. (2006a). The intrinsic forensic discriminatory power of diphthongs. In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 64–69).
Rose, P. (2006b). Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language, 20*(2), 159–191.
Rose, P. (2013). More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech, Language & the Law, 20*(1).
Rose, P., Kinoshita, Y., & Alderman, T. (2006). Realistic extrinsic forensic speaker discrimination with the diphthong /ai/. In *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 329–334).
Rose, P., Osanai, T., & Kinoshitaa, Y. (2003). Strength of forensic speaker identification evidence: multispeaker formant-and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Forensic Linguistics, 10*, 179–202.
Rose, P., & Simmons, A. (1996). F-pattern variability in disguise and over the telephone-comparisons for forensic speaker identification. In *Proceedings of the 6th Australian International Conference on Speech Science and Technology* (pp. 121–126). Canberra: Australian Speech Science and Technology Association.
Ryalls, J., Shaw, H., & Simon, M. (2004). Voice onset time production in older and younger female monozygotic twins. *Folia Phoniatrica et Logopaedica, 56*(3), 165–169.
Sabatier, S. B., Trester, M. R., & Dawson, J. M. (2019). Measurement of the impact of identical twin voices on automatic speaker recognition. *Measurement, 134*, 385–389.
Sambur, M. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech and Signal Processing, 23*(2), 176–182.
San Segundo, E. (2010a). Variación inter- e intralocutor: parámetros acústicos segmentales que caracterizan fonéticamente a tres hermanos. *Interlingüística, 21*, 352–356.
San Segundo, E. (2010b). Parametric representations of the formant trajectories of Spanish vocalic sequences for likelihood-ratio-based forensic voice comparison. *Journal of the Acoustical Society of America, 128*(4), 2394.
San Segundo, E. (2013a). A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application. *Procedia-Social and Behavioral Sciences, 95*, 59–67.
San Segundo, E. (2013b). Guess who is laughing: A perceptual experiment on twin and non-twin siblings' identification. *Paper presented at the 31st International Conference AESLA (Asociación Española de Lingüística Aplicada)*. San Cristóbal de La Laguna: Universidad de La Laguna.
San Segundo, E. (2014). Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics (Doctoral dissertation). CSIC-Universidad Internacional Menéndez Pelayo. Published as monograph in 2017 as Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, Alicante: Biblioteca Virtual Miguel de Cervantes. Retrieved from http://www.cervantesvirtual.com/obra/forensic-speaker-comparison-of-spanish-twins-and-non-twin-siblings-a-phonetic-acoustic-analysis-of-formant-trajectories-in-vocalic-sequences-glottal-source-parameters-and-cepstral-785163/.
San Segundo, E., & Künzel, H. (2015). Automatic speaker recognition of Spanish siblings: (monozygotic and dizygotic) twins and non-twin brothers. *Loquens, 2*(2). https://doi.org/10.3989/loquens.2015.021 e021.
San Segundo, E., & Mompeán, J. A. (2017). A Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity. *Journal of Voice, 1*(5), 644.e11–644.e27. https://doi.org/10.1016/j.jvoice.2017.01.005.
San Segundo, E., Tsanas, A., & Gómez-Vilda, P. (2017). Euclidean distances as measures of speaker dissimilarity including identical twin pairs: a forensic investigation using source and filter voice characteristics. *Forensic Science International, 270*, 25–38. https://doi.org/10.1016/j.forsciint.2016.11.020.
Segal, N. (1984). Cooperation, competition, and altruism within twin sets: A reappraisal. *Ethology and Sociobiology, 5*(3), 163–177.
Segal, N. (1990). The importance of twin studies for individual differences research. *Journal of Counseling & Development, 68*(6), 612–622.
Smits, J., & Monden, C. (2011). Twinning across the developing world. *PLoS One, 6*(9) e25239.

Smith, J., Renshaw, D., & Renshaw, R. (1968). Twins who want to be identified as twins. *Diseases of the Nervous System, 29*(9), 615–618.
Stromswold, K. (2006). Why aren't identical twins linguistically identical? Genetic, prenatal and postnatal factors. *Cognition, 101*(2), 333–384.
van Leeuwen, D., & Brümmer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In C. Müller (Ed.), *Speaker Classification I: Fundamentals, Features, and Methods* (pp. 330–353). Heidelberg: Springer-Verlag.
van Lierde, K., Vinck, B., De Ley, S., Clement, G., & Van Cauwenberge, P. (2005). Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach. *Journal of Voice, 19*(4), 511–518.
Weirich, M. (2011). *The influence of NATURE and NURTURE on speaker-specific parameters in twins' speech: Articulation, acoustics and perception. (Doctoral dissertation)*. Humboldt-Universität zu Berlin.
Weirich, M., & Lancia, L. (2011). Perceived auditory similarity and its acoustic correlates in twins and unrelated speakers. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2118–2121).
Whiteside, S., & Rixon, E. (2001). Speech patterns of monozygotic twins: an acoustic case study of monosyllabic words. *The Phonetician, 82*(2), 9–22.
Whiteside, S. P., & Rixon, E. (2004). Speech characteristics of monozygotic twins and a same-sex sibling: an acoustic case study of coarticulation patterns in read speech. *Phonetica, 60*(4), 273–297.
Wolf, J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America, 51*(6B), 2044–2056.
Yarmey, A., Yarmey, A., Yarmey, M., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology, 15*(3), 283–299.
Zuo, D., & Mok, P. P. K. (2015). Formant dynamics of bilingual identical twins. *Journal of Phonetics, 52*, 1–12.