DOI: 10.17469/O2108AISV000008

ALICE ALBANESI, SONIA CENCESCHI, CHIARA MELUZZI,
ALESSANDRO TRIVILINI

Italian monozygotic twins' speech: a preliminary forensic investigation[1]

In this study, we investigate whether it is possible to distinguish a speaker from his twin in low quality audio recordings. An analysis of both qualitative and quantitative data was conducted to compare the speech of 4 pairs of Italian twins (2F, 2M). The distributions of fundamental frequency and formants were similar across twin pairs, but Lobanov's normalization allowed a differentiation of twins' speech, especially in the elicited form. The statistical analysis confirmed these outcomes and highlighted some differences and the role of F3. The results are discussed in a forensic perspective. Further experiments will widen the sample and the features of interest to determine if this methodology represents a valid procedure for twins' speech discrimination.

Keywords: speaker recognition, monozygotic twins, normalization, formants, between-speaker variability.

1. *Introduction*

In this first exploratory study, a preliminary corpus of monozygotic twins is considered to determine if acoustic cues can discriminate their speech. Twins' speech similarity degree depends on the variable sum of an anatomical inheritance with environmental and social factors, which contribute to define each sibling's personality and behavioural tendencies (Nolan, Oh, 1996).

In forensics, monozygotic twins' voices are an interesting aspect. Although direct involvement of pairs in practical cases is extremely rare, they represent the highest level of physical similarity between two different people and therefore the lowest limit of between-speaker variation (Loakes, 2008; Fernández, 2013). As a consequence, twins are an excellent starting point to study a number of key topics in forensics, such as the acoustic features influencing speaker's recognition accuracy or the auditory discrimination of voices.

When a criminal case involves a pair of monozygotic twins, it must be noted that identifying the offender is far more difficult than normal discrimination tasks

[1] This work has been conceived and written jointly by the four authors. However, for the Italian evaluation system, author 1 is responsible for sections 1, 2, 3, 4.2, and 7 (with author 2); author 2 is also responsible for sections 4, 4.1; author 3 is responsible for sections 5 and 7; author 4 supervised the work. Author 1 performed the recordings and the annotation of the whole corpus, and with Author 2 also conceived the experimental design of the work, while author 3 performed the statistical analysis.

just involving siblings. Indeed, twins' DNA is almost wholly identical; DNA tests can only narrow the field to the pair of siblings but, afterwards, it is impossible to define the single responsible beyond any reasonable doubt (Planterose Jiménez, Liu, Caliebe, Montiel González, Bell, Kayser & Vidaki, 2021; Vidaki, Lopez, Carnero-Montoro, Ralf, Ward, Spector & Kayser, 2017). Recent developments in technological and biological fields have led to new analytical techniques, but despite their ability in making a distinction, they must be evaluated over time (Kader, Gahi & Olaniran, 2020). Hence, this study could also be useful to lay the foundations for a possible recognition system based on other non-genetic features, such as acoustic cues derived from audio recordings.

In this regard, our work proposes a new analytical and comparative method that accounts for social and personal character differences. In light of this general purpose, the following research questions are addressed in this paper:

- Could formants or the fundamental frequency be involved in the discrimination of Italian monozygotic twins' voice in low quality audio recordings?
- Are twins' (dis)similarities reduced or enhanced in controlled vs spontaneous speech?
- Is there a useful methodology to make a distinction by exploiting these features?
- What are the possible implications of these findings for a semi-automatic discrimination of speakers in forensics?

Forensic speakers' comparison aims at evaluating whether the voices extracted from two audio recordings belong to the same person. The present study is designed to identify all those aspects that may be consistent with differences between speakers, rather than a match between their voices. If the results show the presence of such differences, the forensic consequence would be the introduction of a new analytical and comparative method in the investigative field, capable of simplifying inquiries concerning monozygotic twins, as well as useful for studying siblings' sociophonological issues.

The paper is organised as follows: section 2 presents the theoretical premises of the work, while section 3 describes the corpus with the associated technical issues. Section 4 and 5 present the outcomes of both a qualitative and a quantitative analysis on formants' variability. Finally, section 6 discusses the results and possible applications in forensic phonetics, and section 7 reports our conclusions and further perspectives.

2. *Theoretical premises*

In this study, we assume monozygotic twins represent the most extreme physical similarity between two subjects, and consequently the lowest possible variation between speakers (Fernández, 2013).

On the one hand, the speaker's characteristic timbre depends on the biological conformation of his resonator, which is similar in monozygotic twins. On the other side, prosodic variability could influence timbre perceptions through many

different acoustic cues (Berry, Brown 2019, and see also Jensen 2002 on timbre and other involved acoustic cues). Additionally, prosody is affected by several aspects including regional and social accents, as well as emotion, context, and relational skills (Cenceschi, Sbattella & Tedesco, 2018).

Anatomical inheritance, as well as environmental and social factors, intertwine in determining the similarity degree within twins' speech (Fernández, 2013). Given that genetic heritage matches, we can hypothesize that vocal timbre, also defined as the starting point, is supposedly remarkably similar at a young age in twin pairs, but it may evolve and change with growth (Gahl, Baayen, 2019), allowing a possible distinction of the single sibling's voice.

Although a rich international scientific literature on twins' speech exists (van Braak, Heeren, 2015; Johnson, Azara, 2000; Sebastian, 2013; Zuo, Mok, 2013; Van, Vercammen & Debruyne, 2001), there are far fewer studies addressing Romance languages. San Segundo et al. (2017) exploited the Euclidean Distances (considering F0 and MFCCs) to investigate the similarity of Spanish voices between different speakers, same speaker and twin pairs, in both high quality and telephone-filtered recordings. Besides presenting the intuitive result that twins similarity lies between values for the first and second case, this study portrays an exhaustive synthesis of the scientific works attempting to differentiate monozygotic siblings' speech.

In spite of the presence of several Italian speech corpora (Falcone, Gallo, 1996; Cresti, Moneglia, 2005), nothing seems to exist for twins' voices. Few investigations focus on the phonetic differences between twins' speech based on perceptual acoustic cues (e.g., Giannini, 1989; Gedda, Bianchi & Bianchi-Neroni, 1958). However, the results are still limited by the small sample size and by the fact that they do not deal with forensic issues. Moreover, to the authors' knowledge, ASRs usually focus on general speakers' discrimination, but it is not clear how they would react to the voices of monozygotic siblings. Contrastingly, Künzel's (2010) ASR evaluated an earlier GMM/UBM system and discovered that it could work under certain conditions, but its performance is negatively impacted by the genetic similarity, especially for spontaneous speech, different amounts of audio for each speaker, and for female twins. San Segundo, Künzel (2015) tried a similar test with BatvoxTM reporting comparable findings, and suggesting that the cepstral parameters the automatic system BatvoxTM is based on are genetically influenced.

3. *Method and materials*

This preliminary study focuses on a small corpus consisting of 4 pairs of monozygotic twins (4 males and 4 females), aged between 20 and 25 years old, with the same sociolinguistic characteristics, such as high level of education, born and living in the north-west of Italy, L1 Italian. All speakers consciously and freely took part in the experiment as volunteers, receiving no compensation for their participation in the project.

Vocal data collection followed the Interactive Atlas of Romance Intonation guidelines (Prieto, Borràs-Comes & Roseano, 2010-2014), which ensured vocal tasks with a mixed prosodic model, already widely assessed in sociolinguistic and prosody-focused research (Frota, Prieto 2015, among others).

Each speaker performed two different tasks guided by the first author, in order to simulate two opposite emotional states: a stressful condition and a comfortable (relaxed) one.

The first task concerned the elicited speech: people were introduced to a list of 31 questions regarding different contexts and situations, asking them to answer as fast as they could and in a coherent way. Every speaker received the same kind of questions in the same order, allowing answers that were as much similar as possible in terms of content and intonation. Afterwards, twin pairs had to undergo the second task: a short interview lasting an average of 7 minutes in order to acquire spontaneous speech samples. Questions concerned general topics of daily life to make them feel at ease and facilitate an instinctive response.

3.1 Audio recording modality

Each twin was recorded alone to avoid accommodation phenomena (either convergent or divergent) or mutual emotional influences. Furthermore, all recordings were performed by the first author to limit the possible convergence effect towards other speakers. Vocal samples were collected with a smartphone recording app (see § 2.2) in a silent room to avoid particularly intense or continuous background noise. However, due to Covid-19 pandemic restrictions, only the first two pairs of siblings (named M1-M2 and F1-F2) were recorded through a face-to-face interview. The third and fourth couple's (named M3-M4 and F3-F4) audio files were gathered via phone call, where the speaker used one device to communicate with the author and a second device to record the vocal samples.

Data were stored according to tasks, and speakers were identified by an alphanumeric code: F for Female, M for Male, and an increasing number from 1 to 4 according to the recording order.

3.2 Technical equipment and digital formats

In the forensic context, audio files are often collected with different electronic devices and are typically characterized by an exceptionally low quality, due to hard and noisy recording conditions, narrow frequency bands and compressed formats (Cenceschi, Meluzzi & Nese, 2020). In order to simulate a classical forensic setting, our samples have been acquired by using smartphones of assorted brands and models. Each pair of siblings used a single device to provide vocal data: iPhone-6 and Samsung Galaxy S10, Galaxy S9, Galaxy S6. These devices created mp3 (44.1 kHz – 16 bit) or m4a mono files (48-44 kHz – 16 bit) which were later converted into wav files (44.1 kHz – 16 bit) to enable subsequent elaborations with Audacity 2.3.3 and PRAAT 6.0.52 (Boersma, Weenik, 2019). As a following step, files were

separated with Audacity and stored for task, obtaining the current wording: number of sentences followed by the speaker's code (e.g., 1 M1.wav).

3.3 Features' extraction

The analysis focused on three target vowels /a/, /e/, /o/. For each sentence, they were manually annotated with PRAAT TextGrid option, considering the left and right boundaries of each vowel as the beginning and the end of the second formant (F2), and rejecting all vowels that were not clearly discernible at listening. Aiming to emulate truthful forensic conditions (short or partially unusable recordings without possibility to compare target phonemes), all possible vowels were tagged (Rhodes, 2012) obtaining the overall corpus as shown in Tab. 1.

As a result, we could observe vowels' overall variability depending on the production modality (elicited speech or interview mode).

Once the annotation was concluded, using a PRAAT script, we automatically extracted the following acoustic parameters at the midpoint of the target vowels: F0, F1, F2 and F3. Formant values were manually corrected to remove outliers, and visually inspected through web application Visible Vowels (Heeringa, Van de Velde, 2017). Finally, we created a second parallel table with formant values normalized through Lobanov z-transformation to Hertz value (Lobanov 1971, see also van der Harst 2011: 97), that is without a previous conversion of Hertz values into Bark (as tried, for instance, by Rietveld, van Houven 2009). As summarized by Adank, Smits & van Hout (2004), Lobanov's formula represents a formant-intrinsic and vowel-extrinsic transformation, which is claimed to better preserve sociophonetic variability by minimizing the possible between-speaker biological differences (cf. van der Harst 2011: 315).

Table 1 - *The total amount of tokens per vowels and communicative task*

	Total vowels (8 speakers)	
Vowel	**Elicited sentences**	**Sponteneous speech**
A	1006	2000
E	986	1710
O	769	1526

Data were analysed both qualitatively and quantitatively. This was done for the potential forensic applications of the findings, and also in order to highlight possible biases in both kinds of analysis. As for the quantitative evaluation, the main purpose was to assess which acoustic cues perform better at distinguishing siblings in elicited and spontaneous speech, and whether there are differences between the statistical results and the qualitative ones.

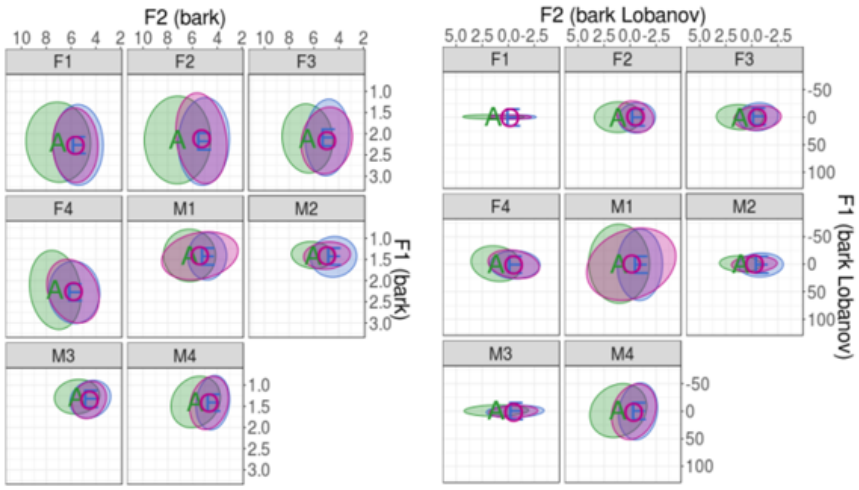
4. *Qualitative analysis*

Nowadays, there are no forensic standard thresholds to assess the level of (dis) similarity between speakers' formants, even less for Italian and identical twins, and the evaluation is committed to the single analyst experience. Therefore, in the present study, we provided a preliminary analysis based on vowels' distribution crossed with social information. Through Visible Vowels, as suggested in Cenceschi, Meluzzi & Trivilini (2021), we represented the vowel space for F1-F2 and F2-F3 of the complete set of target vowels (/a/, /e/, /o/) for each speaker, applying different units and normalization methodologies.

It should be noted that we cannot expect a formants' behavior in line with the standard values for Italian vowels as highlighted by Giannini, Pettorino (1992), because we are analyzing vowels extrapolated from a smartphone recording app in order to imitate the typical conditions of real forensic cases. Therefore, alterations due to the codec used for compression, different physical distances of the speakers from the microphone, and different environmental recording conditions are proven (although the environments were not noisy, a control on this variable was not intentionally introduced) as reported, among others, in Cheng, Burnett (2011), Byrne, Foulkes (2004), and Künzel (2001,2002). Moreover, lower bitrates cause a decrease in the reliability of feature values, introducing important alterations compared to high quality audio formats (Gonzalez, Cervera & Llau, 2003).

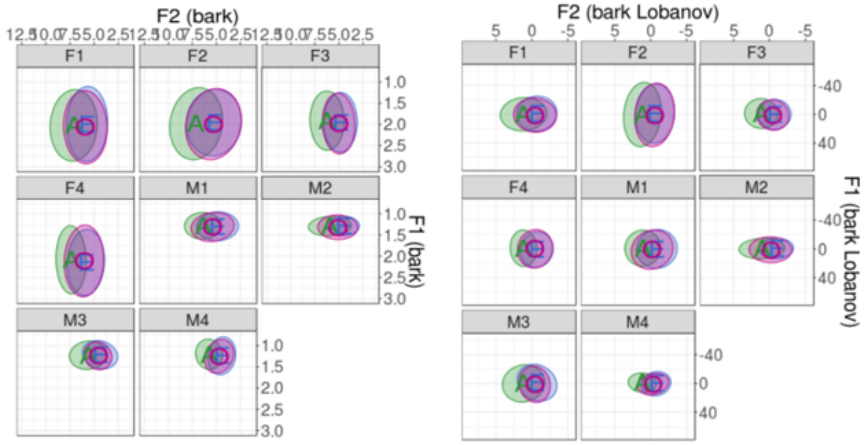
In particular, it may be noted that all the F1-F2 vowel spaces reflect what already tested by Byrne, Foulkes (2004) and Künzel (2001) regarding the non-linear behavior (across different vowels) of the first two formants in critical conditions for quality and recording, and their high variability with different tokens. For example, closed vowels show a greater F1 raising effect than more open vowels, and the F2 of closed front vowels may be subjected to a significant lowering. Moreover, the necessity to analyze as many vowels as possible, due to the lack of materials, stresses even more the need for caution in drawing forensic conclusions (Byrne, Foulkes, 2004). In light of this premises, vowel spaces do not show the exact values of the single vowel's formants, but rather their global dispersion, which can be useful for extrapolating qualitative information to add to the forensic report (see also Cenceschi, Meluzzi & Trivilini, 2021).

Figure 1 - *Vowel space for elicited speech before and after Lobanov's normalization*



The analysis of the third formant is not included in the current study since its diagrams do not show substantial differences even after normalization, probably because it is more linked to the vocal tract length, which is extremely similar in monozygotic twins. We started by analyzing the un-normalized vowel areas for elicited speech followed by spontaneous speech, comparing twins of the same pair. Afterwards, we considered the graphics obtained with Lobanov's normalization (Flynn, Foulkes, 2011) as shown in Fig. 1 and 2, highlighting the differences detected before and after the transformation for each speaker, as well as those existing with their co-twin.

Figure 2 - *Vowel space for spontaneous speech and for all speakers before and after Lobanov's normalization*



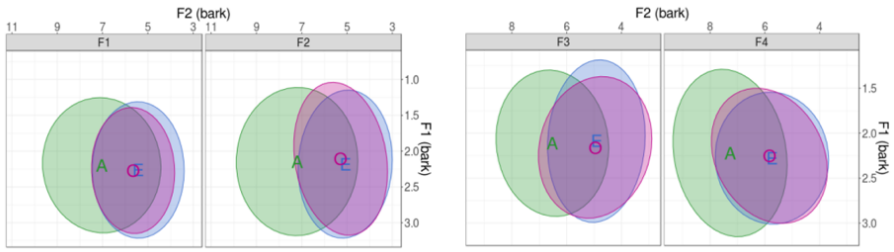
4.1 Analysis of un-normalized data

4.1.1 Comparison between twins belonging to the same pair

We compared the vowel space of each couple in controlled speech (sentences) considering values of F0 and un-normalized formants. We noticed a strong level of correspondence between the two siblings' diagrams both in Hertz and Bark units (Traunmüller, 1990).

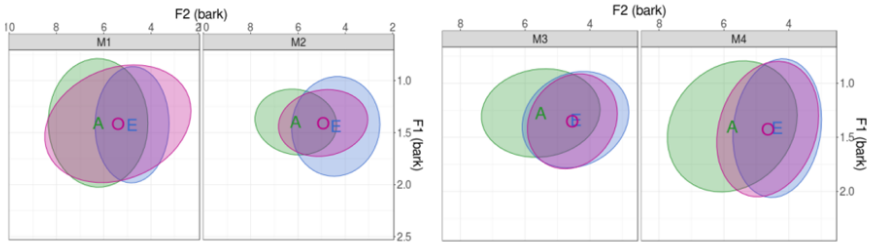
Each target vowel shows the same centre of gravity and a remarkably similar distribution of formants. As displayed in Fig. 3, the distributions of the two female pairs (F1-F2 and F3-F4) are pretty resembling.

Figure 3 - *Female couples' vowel space*



Additionally, there is a larger distinction in the vowel space of /a/ and /o/ in the first male couple M1-M2 (wider distribution for M1), while the centre of gravity location is the same for both brothers. Indeed, an almost total correspondence is detected in the properties of /e/ (Fig. 5). Similarly, the other pair of twins M3-M4 reveals a slight difference in distributions and in the midpoint of /a/, a little higher in M3 (Fig. 4).

Figure 4 - *Male couples' vowel space*



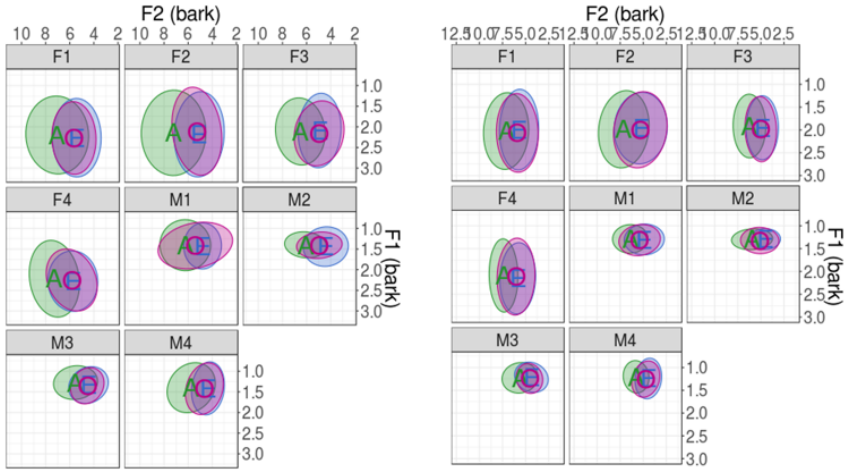
However, it must be noted that the same level of between-vowel-space dissimilarity could also be found comparing the speech of a single person recorded in disparate moments (e.g., different social context or emotional state).

Moreover, male speakers show an intra-pair difference in the vowel area distribution which is a little more evident than for female couples, but the sample is too small to allow further generalizations.

4.1.2 Comparison between elicited and spontaneous speech for the same speaker

When considering values of F0 and un-normalized formants for spontaneous speech, each speaker exhibits a smaller vocal area than in elicited sentences. Moreover, the centres of gravity tend to overlap (Fig. 5). This will be the focus of interesting considerations we shall address later.

Figure 5 - *Elicited and spontaneous speech for all pairs*



4.1.3 Comparison between spontaneous speech of twins belonging to the same pair

Spontaneous speech heightens the similarity between siblings of the same female couples (Fig. 6) with respect to the elicited form (Fig. 7).

Figure 6 - *Spontaneous speech vowel space for female pairs*

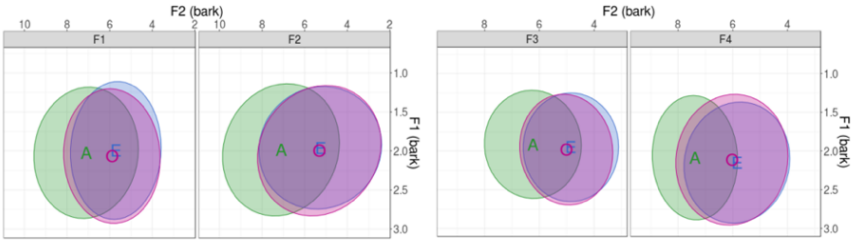
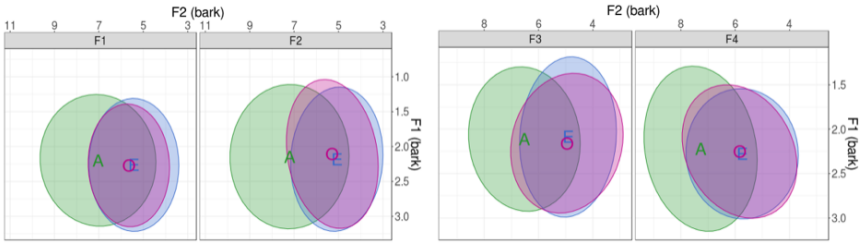


Figure 7 - *Elicited speech vowel space for female pairs*



The same pattern was observed for male speakers, as shown in Fig. 8 and 9. However, compared to the other pairs, M3 and M4 display a slightly greater difference, even though still considered irrelevant.

Figure 8 - *Spontaneous speech vowel space for male couples*

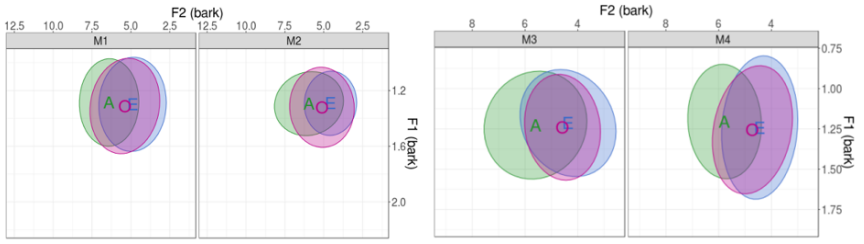
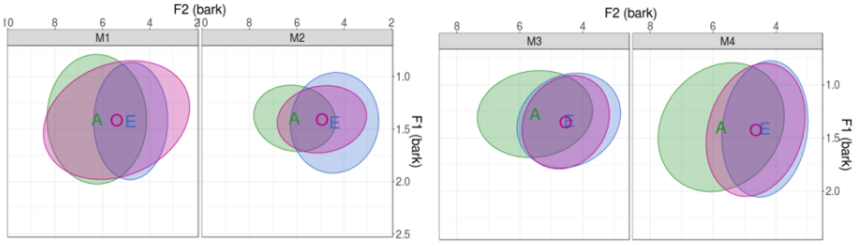


Figure 9 - *Elicited speech vowel space for male couples*



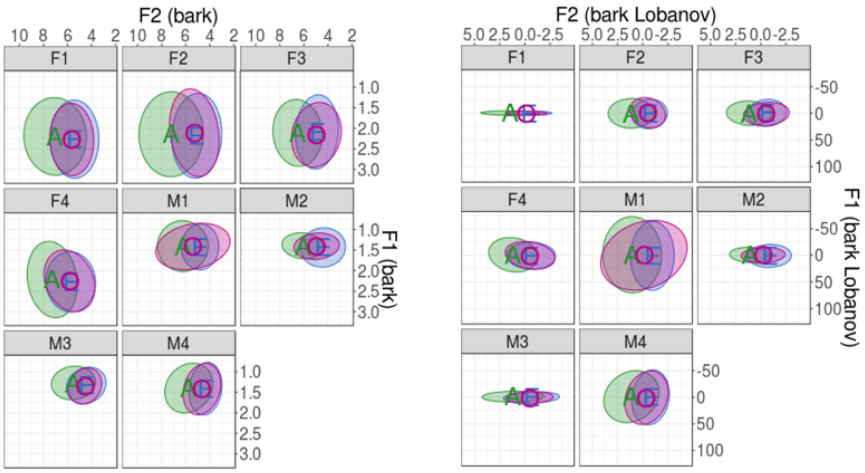
4.2 Analysis of normalized data

In order to look for more significant results, we opted for a normalization of formants with Lobanov's formula. According to many studies in sociophonetics (e.g., Van der Harst, 2011), this procedure is preferred, since it reduces the difference between formantic values due to physiological factors and preserves more information about sociolinguistic variants. It would underline the dichotomies coming from social and environmental influence, emotional perturbations, personality, and the context of discussion the individual is facing during the verbal production (Van der Harst, 2011; Adank, 2003). Although this normalization usually works better with huge corpora, we proceeded anyway to understand its possible usefulness in a forensic

context, where recordings are often short in duration, and speakers to be compared are quite always limited in number.

Confronting the vowel space of each person before and after Lobanov's normalization, we observed that M1, F2 and M4 kept on being almost identical, unlike what happened to their co-twins whose acoustic representations were quite different. Only the pair F3-F4 did not undergo variations for both speakers (Fig. 10).

Figure 10 - *Vowel space before and after Lobanov's normalization*



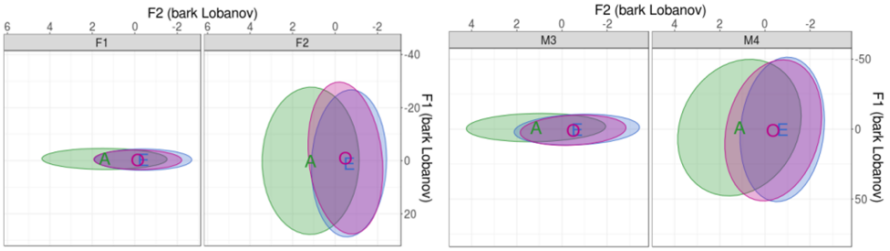
4.2.1 Comparison of co-twins' normalized data in elicited speech

Before normalization, vowel areas looked remarkably similar between co-twins, but once the procedure was applied, substantial differences could be observed in 3 out of 4 pairs.

Comparing normalized data of twins belonging to the same couple (e.g., Fig. 11), we noticed that the second formant maintains a similar distribution between siblings for all vowels. According to empirical theories, the more stable frequency (F2) is less affected by factors unrelated to the individual's physiology (Adank, 2003).

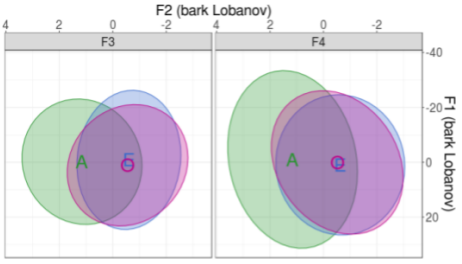
The first formant's behaviour, on the other hand, is extremely different and enables a speaker discrimination; this would confirm the theory of Adank et al. (2004) in attributing to Lobanov's normalization the ability to highlight non-physiological but sociolinguistic dichotomies. A comparison of these distributions also suggested that twins exploit different vowel spaces differently: one employs anteriority-posteriority, while the other exploits height.

Figure 11 - *Comparison of normalized distributions of a female and a male couple*



The only exception is represented by the F3-F4 couple. In this case, we observed smaller dissimilarities among distributions (Fig. 12).

Figure 12 - *Comparison of normalized distributions for the second female couple*



4.2.2 Comparison of co-twins' normalized data in spontaneous speech

Evaluating the spontaneous speech, we observed comparable results to those obtained with elicited tasks. In particular, the first formant continues to be the parameter that, when normalized, enables a distinction between speakers (Fig. 13). However, this phenomenon is less pronounced than in the elicited speech (Fig. 14).

Figure 13 - *Un-normalized Vs normalized vowel space for spontaneous speech*

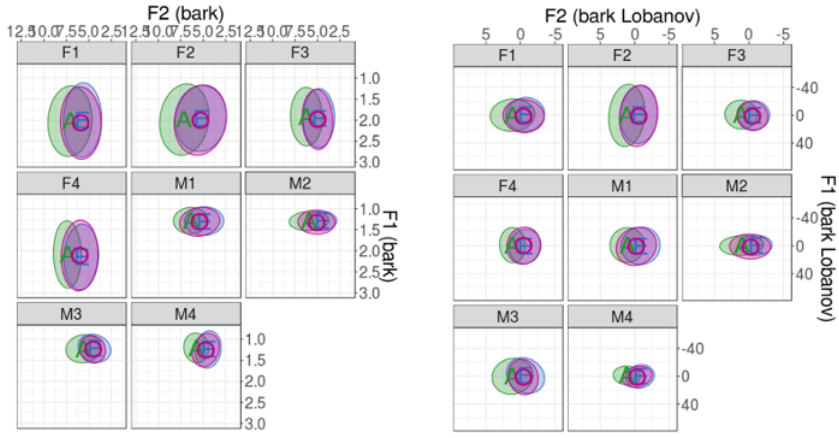
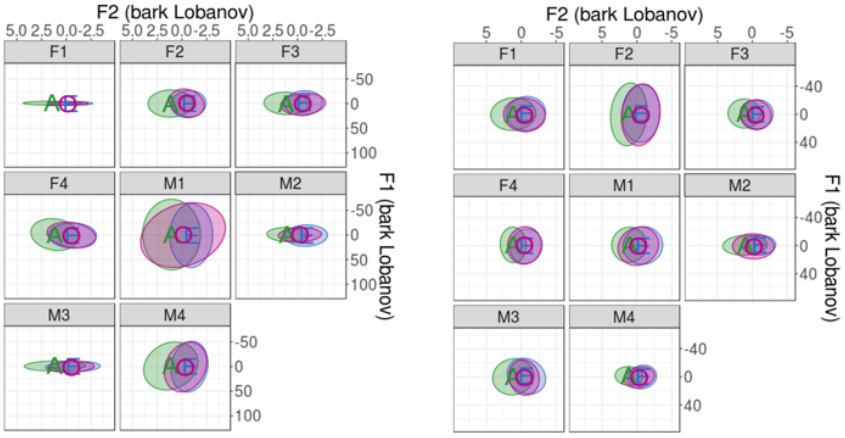
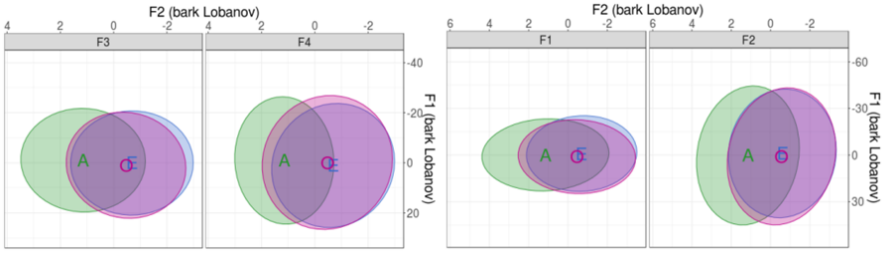


Figure 14 - *Elicited Vs Spontaneous normalized vowel space*



F3 and F4 are an exception as their distributions remain similar (Fig. 15). However, there are some slight discrepancies in all the other pairs, whose diagrams manifest a disparity in the shape and width of vowel areas.

Figure 15 - *Normalized distributions for the second female couple*



5. *Statistical analysis*

We performed different ANOVAs for vowels /a/, /e/ and /o/ on each twin couple, by applying Bonferroni correction for splitted datasets. The main purpose of this analysis was to define whether the mean differences in formantic values between siblings of each pair were statistically relevant, that is whether a statistical method commonly used in forensics (when statistics is applied, that must be said) could be helpful to discern similar voices. The analysis has been performed on IBM SPSS 21 for F0, F1, F2 and F3. Tab. 2 sums up the results obtained for the whole corpus.

Table 2 - *The results of the ANOVAs for vowel quality and group; significant correlations have been highlighted*

	Group 1 (F1-F2)			**Group 2 (F3-F4)**		
	/a/	**/e/**	**/o/**	**/a/**	**/e/**	**/o/**
F0	F(702)=1.045; p=.307	F(642)=.599; p=.439	F(502)=1.665; p=.198	F(802)=23.363; p<.0001	F(664)=40.196; p<.0001	F(520)=12.432; p<.0001
F1	F(702)=1.040; p=.308	F(642)=121.103; p<.0001	F(502)=24.206; p<.0001	F(802)=297.087; p<.0001	F(664)=231.543; p<.0001	F(520)=162.941; p<.0001
F2	F(702)=2.455; p=.118	F(642)=19.349; p<.0001	F(502)=.549; p=.459	F(802)=.329; p=.566	F(664)=8.750; p=.003	F(520)=1.287; p=.257
F3	F(702)=59.601; p<.0001	F(642)=10.625; p<.0001	F(502)=25.583; p<.0001	F(802)=116.499; p<.0001	F(664)=229.77; p<.0001	F(520)=93.901; p<.0001
	Group 3 (M1-M2)			**Group 4 (M3-M4)**		
	/a/	**/e/**	**/o/**	**/a/**	**/e/**	**/o/**
F0	F(835)=1.872; p=.172	F(800)=2.073; p=.150	F(746)=6.084; p=.014	F(659)=17.851; p<.0001	F(659)=21.418; p<.0001	F(520)=10.191, p<.0001
F1	F(835)=3.019; p<.0001	F(800)=26.631; p<.0001	F(746)=8.407; p=.004	F(659)=10.931; p<.0001	F(659)=.113; p=.737	F(520)=4.80; p=.029
F2	F(835)=2.361; p=.125	F(800)=2.947; p=.086	F(746)=.998; p=.318	F(659)=29.825; p<.0001	F(659)=11.852; p<.0001	F(520)=.118; p=.731
F3	F(835)=137.26; p<.0001	F(800)=75.497; p<.0001	F(746)=29.602; p<.0001	F(659)=.212; p=.645	F(659)=15.473; p<.0001	F(520)=1.081; p=.299

Data showed that the third formant seemed to be more sensitive in discriminating siblings, since the only exception is represented by the M3-M4 pair, where only vowel /e/ showed a significant difference between the two speakers. The second formant was relevant just for front vowels and for /e/ in particular, whereas the situation was more scattered for the first formant, as well as for the fundamental frequency. It also seemed that discrepancies highlighted for the first female pair (F1-F2) were in general less meaningful than in other pairs of our sample. The same could be said for the first male couple (group 3) when compared to the second one (group 4). However, the situation appears slightly different when we check dissimilarities by also dividing for speech task, which is elicited (Tab. 3) and spontaneous (Tab. 4) speech.

Table 3 - *The results of the ANOVAs for vowel quality and group in elicited speech; significant correlations have been highlighted*

	Group 1 (F1-F2)			**Group 2 (F3-F4)**		
	/a/	/e/	/o/	/a/	/e/	/o/
F0	F(231)=1.407; p=.237	F(256)=3.574; p=.060	F(163)=1.229; p=.269	F(291)=10.023; p=.002	F(279)=7.667; p=.006	F(215)=4.683; p=.032
F1	F(231)=6.602; p=.011	F(256)=3.903; p=.049	F(163)=7.451; p=.007	F(291)=46.677; p<.001	F(279)=74.465; p<.001	F(215)=39.751; p<.001
F2	F(231)=.239; p=.626	F(256)=.184; p<.001	F(163)=.965; p=.327	F(291)=5.275; p=.022	F(279)=5.112; p=.025	F(215)=5.899; p=.016
F3	F(231)=40.228; p<.001	F(256)=.1.478; p=.225	F(163)=23.488; p<.001	F(291)=24.151; p<.001	F(279)=67.611; p=.001	F(215)=13.539; p<.001

	Group 3 (M1-M2)			Group 4 (M3-M4)		
	/a/	/e/	/o/	/a/	/e/	/o/
F0	F(248)=.249; p=.619	F(221)=1. 265; p=.262	F(200)=.557; p=.456	F(228)=23. 998; p<.001	F(225)=9. 206; p=.003	F(183)=13. 469; p<.001
F1	F(248)=5. 786; p=.017	F(221)=14. 168: p<.001	F(200)=5. 055; p=.026	F(228)=9. 100; p=.003	F(225)=.336; p=.563	F(183)=.752; p=.387
F2	F(248)=5. 614; p=.019	F(221)=8. 095; p=.005	F(200)=.014; p=.906	F(228)=23. 921; p<.001	F(225)=14. 650; p<.001	F(183)=.068; p=.794
F3	F(248)=49. 140; p<.001	F(221)=.393; p=.531	F(200)=1. 409; p=.237	F(228)=3. 086; p=.080	F(225)=14. 762; p<.001	F(183)=.203; p=.653

Table 4 - *The results of the ANOVAs divided for vowel quality and group in spontaneous speech; significant correlations have been highlighted*

	Group 1 (F1-F2)			Group 2 (F3-F4)		
	/a/	/e/	/o/	/a/	/e/	/o/
F0	F(469)=.354; p=.552	F(384)=.808; p=.369	F(337)=.811; p=.369	F(509)=23. 170; p<.001	F(383)=45. 211; p<.001	F(303)=12. 342; p<.001
F1	F(469)=..629; p=.428	F(384)=20. 753; p=.001	F(337)=16. 530; p=.001	F(509)=323. 340; p<.001	F(383)=159. 102; p<.001	F(303)=127. 689; p<.001
F2	F(469)=2. 168; p=.142	F(384)=34. 303; p=.001	F(337)=.062; p=.803	F(509)=5. 557; p=.019	F(383)=4. 378; p=.037	F(303)=.025; p=.874
F3	F(469)=26. 252; p=.001	F(384)=9. 357; p=.002	F(337)=9. 447; p=.002	F(509)=104. 612; p<.001	F(383)=166. 502; p<.001	F(303)=94. 601; p<.001
	Group 3 (M1-M2)			**Group 4 (M3-M4)**		
	/a/	/e/	/o/	/a/	/e/	/o/
F0	F(585)=1. 114; p=.292	F(577)=5. 050; p=.025	F(544)=1. 029; p=.311	F(429)=3. 636; p=.057	F(359)=13. 604; p<.001	F(335)=.910; p=341
F1	F(585)=29. 143; p<.001	F(577)=8. 445; p=.004	F(544)=2. 433; p=.119	F(429)=3. 247; p=.072	F(359)=.812; p=.368	F(335)=4. 560; p=.033
F2	F(585)=.237; p=.626	F(577)=11. 544; p<.001	F(544)=1. 154; p=.283	F(429)=10- 154; p=.002	F(359)=2. 265; p=.133	F(335)=.047; p=.829
F3	F(585)=101. 077; p<.001	F(577)=125. 317; p<.001	F(544)=38. 478; p<.001	F(429)=.830; p=.363	F(359)=2. 772; p=.097	F(335)=2. 682; p=.102

From the ANOVAs reported in Tab. 3 and 4, discrepancies between siblings' formantic values are more evident in elicited than in spontaneous speech. However, this is also variable across groups. Indeed, the second female group (speakers F3 and F4) showed in general more statistically relevant variations if compared to all the other groups. Among male subjects, M3-M4 differ significantly in elicited speech with respect to vowels /a/ and /e/, but these differences are almost non-existent in the spontaneous form. Conversely, M1-M2 and F1-F2 behave more similarly across tasks. In general, these data confirm that F3 is good at predicting variability among twins' speech (with the notable exception of the fourth group in spontaneous speech). The first formant, as well as the front vowel /e/ and the central vowel /a/, also performed well at discriminating between siblings within the same group.

6. *Discussion*

To discuss potential solutions for monozygotic twins' acoustic distinction, we proceeded by synthesizing the obtained results and crossing them with social and context information:

- F0 and F3 distributions were extremely similar for co-twins and did not undergo major variations (especially if normalized); as for F0 this has been statistically confirmed, while for F3 the statistical analysis has shown its reliability as a quantitative cue for an intra-twin discrimination.
- Vowel spaces and distributions of the first two formants were extremely similar for co-twins within the same group.
- The only normalization that brings out evident differences between co-twins' speech, for the first formant, is through Lobanov's formula.
- Lobanov's normalization seems to be more effective for elicited speech than for spontaneous speech, although in both cases it brings out variations between co-twins.
- The different discriminatory value of the tasks has also been confirmed by the statistical analysis, although it seems to be strongly group dependent (see below).

The most interesting parameter is therefore the first formant, when data have been subjected to Lobanov's transformation. Assuming that Lobanov evidence could underline socio-phonetic differences related to the context of growth (Van der Harst, 2011; Adank, 2003), we tried to sketch some preliminary hypothesis and logical remarks based on the personal history of speakers.

The first and fourth couple (M1-M2; F3-F4) meet predictions and support this theory. In spite of living together and pursuing university degrees, the male twins embrace vastly different social circumstances, such as their university path, sports, employment and love lives.

In fact, M2 shows more stable formants, with a more contained extension of the vowel area linked to the lower F1 variation, while M1 presents opposite characteristics.

On the contrary, F3 and F4 have many things in common and lead similar lives. They live together, attend the same academic year, share a passion for the same sport they have been practicing together for years, and neither of them is romantically involved. The close similarity in social life matches with a strong resemblance in their vowel space.

However, when tested quantitatively, data from Tab. 3 and 4 revealed that F3 and F4, although visually very similar in terms of vowel space, displayed a statistically significant variability in the mean values of all formants (with the exception of the second formant of /o/ in spontaneous speech). This led to two distinct kinds of considerations: the first one concerns the impact of normalization procedures, and the second one focuses on the interaction between qualitative and quantitative analysis for forensic purposes (see 6.1).

As we have seen, compared to the other pairs in our study, F3-F4 twins were almost undistinguishable after Lobanov's normalization. Instead, for both F1-F2 and M3-M4, even though the social context matched, Lobanov underlined some differences in their formants' variability.

Hence, we concluded that Lobanov normalized results could be explained by considering speakers' emotional sphere, which is influenced by their life background. Social conditions and life experiences of twins could be similar, but it does not imply (and it is unrealistic to think so) that their emotional responses and subjective characters are also aligned. In this scenario, twins may have a different emotional response to the elicited task (for instance, nervous or extroverted), resulting in different vowel areas with Lobanov's formula.

Supporting this hypothesis, we noticed that subjects who were more nervous during the formulation of the 31 elicited sentences showed a greater variability of the first formant than their co-twin, who seemed to be calmer and more relaxed. For example:

- The recording of the F2 female has been repeated several times, since her fear of making mistakes induced her to a nervous laughter and frequent interruptions. Conversely, her sister, speaking in confidence with the technician (the first author), showed determination and firmness throughout the acquisition.
- M1 was asked to repeat the test, livening it up to sound less monotonous; therefore, the tension coming from the need to follow precise directions could have been a determining factor. His brother M2, on the other hand, adopted a less variable prosody, reciting the sentences with less emphasis and less marked prosodic variability; not having to meet specific requests, he seemed to be calmer and more relaxed as he was not subjected to pressure.
- This contrast between un-normalized and normalized vowel spaces was also found in the third pair, who presented less variable formantic values in M3 than in M4. As we acquired the audio samples through a telephone call, we were unable to assess the subjects' state of mind, but it is plausible that they approached the task with a different emotional attitude, and that normalization accentuated their conflicting emotions.
- F3 and F4 speakers were apparently both noticeably quiet during the test, and marked differences in temperament could not be discerned. Since both of them have similar lives and social relationships, it is possible that normalization was not able to differentiate them at an emotional level.

6.1 Application in Forensics

The potential impact for this (extremely preliminary) study in the forensic field is two-fold. As addressed in the previous section, there was not always a straight correspondence between the results of the visual (qualitative) analysis of vowel space and the statistical analysis as performed on different formants. The main distinction has been highlighted for the second female pair (F3-F4), whose vowels' distribution appeared similar, especially after Lobanov's transformation, but in the statistical research their formants' values were the most different among the groups of our study. This would suggest a need to be cautious when preferring one analysis to the other. Due to the limited amount of data to investigate, qualitative examination is sometimes preferred in forensics, especially since its results are more understandable by relevant

third parties (e.g., judges and lawyers). However, it remains to be investigated whether formants' differences highlighted by the quantitative analysis have a real phonetic counterpart, which is whether these dichotomies are perceivable by listeners. A perceptual experiment, carefully designed and balanced, may help clarify these (only apparent) contradictions between qualitative and quantitative outcomes.

Furthermore, this work suggests a possible starting point for cases of uncertain attribution through the recording of a phonic test for each twin. Performing a test that "puts under emotional pressure" or causes different speech moods in the two speakers could allow to enhance differences in vowel spaces and in the first formant's distribution with Lobanov's normalization.

As an example, in the Italian authority, phonic tests are normally recorded by asking suspected people to perform a spontaneous speech task within some lists of target words or sentences. On the contrary, if the conclusions derived from the present examination will be validated in further studies, elicited/controlled speech will be more important than the spontaneous one for twins' discrimination.

We state that considerations exposed so far emerged for the first time while performing the present research, therefore it is still not possible to know their value. For this reason, further work will involve a greater number of twin pairs, and we intend to integrate our analysis with other acoustic parameters such as jitter, shimmer and MFCCs. Moreover, vowel observations will be refined in order to study in detail the proximity to certain consonants and their dynamic behaviour (over 7-time steps). A parallel investigation concerns the perceptual aspect: once the analysis of the corpus is completed, a perceptual test will be prepared to understand whether the hearing system validates the results. The hope is that this preliminary outcome could lead to innovative solutions for monozygotic twins' distinction, but also to perform in-depth analysis on sociophonetic variability.

7. *Conclusions*

Drawing a preliminary conclusion, the fundamental frequency, the third formant and the first two un-normalized formants of our monozygotic twins have too similar values and do not allow a visual discrimination (both in Hz and Bark). Despite the work being focused on a limited number of speakers and sentences, it is entirely possible that this finding could be extended to any monozygotic pair living in similar social conditions and sharing past backgrounds. Lobanov's formula seems to be a suitable method to be explored, especially for what concerns elicited speech or, better, recording in stressful conditions.

Lobanov's normalization of spontaneous speech formants always highlights discrepancies between speakers of each pair, but less than in the elicited form. However, this provides a valid contribution supporting the hypothesis that Lobanov's transformation could highlight emotional and attitudinal differences in siblings with the same genetic makeup and aligned life experiences. In the spontaneous speech task, the speaker was free to express himself without interruptions by the

technician. This suggests that people found themselves in a more comfortable situation, with the consequence that different emotional characters did not come to light. Moreover, we must also take into account that spontaneous speech could emphasize more the phenomenon of accommodation, which surely occurs in subjects who are in daily contact and who live in the same families, just like our speakers. Then, accommodation could be mitigated by the executive restrictions of the elicited task, where emotions seem to predominate.

Bibliography

Adank, P., Smits R. & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. In *Journal of the Acoustical Society of America*, 116(5), 3099-3107.

Berry, M., Brown, S. (2019). Acting in action: Prosodic Analysis of Character Portrayal during Acting. In *Journal of Experimental Psychology*, 48(8), 1407-1425.

Byrne, C., & Foulkes, P. (2004). The 'mobile phone effect' on vowel formants. In *The International Journal of Speech, Language and the Law*, *11*(1), 83-102.

Boersma, P., Weenink, D. (2019). PRAAT: doing phonetics by computer [Computer program]

Version 6.0.52, retrieved from http://www.PRAAT.org/

Cenceschi, S., Meluzzi, C. & Trivilini, A. (2021). The Variability of Vowels' Formants in Forensic Speech. In *IEEE Instrumentation & Measurement Magazine*, 24(1), 38-41.

Cenceschi, S., Meluzzi, C. & Nese, N. (2020). Speaker's identification across recording modalities: a preliminary phonetic experiment. In Romito L. (Ed.). *Language change under contact conditions: acquisitional contexts, languages, dialects and minorities in italy and around the world,* Studi AISV 7. Milano: Officinaventuno, 407-426.

Cenceschi, S., Sbattella, L. & Tedesco, R. (2018). Verso il riconoscimento automatico della prosodia. In Bertini C., Celata C., Lenoci G., Meluzzi C & Ricci I. (Eds.). *Social and biological factors in speech variation,* Studi AISV 3. Milano: Officinaventuno, 433-440.

Cheng, E., & Burnett, I.S. (2011). On the effect of AMR and AMR-WB GSM compression on overlapped speech for forensic analysis. In *Proceeding of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1872-1875.

Cresti, E., Moneglia, M. (Eds.). (2005). *C-oral-rom: integrated reference corpora for spoken romance languages*. Amsterdam: John Benjamins publishing.

Falcone, M., Gallo, A. (1996). The "SIVA" speech database for speaker verification: description and evaluation. In Bunnell, H.T., Idsardi, W. (Eds.) (1996). *Proceeding of Fourth International Conference on Spoken Language Processing,* Philadelphia, USA, 3-6 October 1996, 1902-1905.

Fernández, E.S.S. (2013). A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application. In *Procedia-Social and Behavioral Sciences*, 95, 59-67.

Frota, S., Prieto, P. (Eds.). (2015). *Intonation in Romance*. Oxford: Oxford University Press.

Gahl, S., Baayen, R.H. (2019). Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood. In *Journal of Phonetics*, 74, 42-54.

Giannini, A. (1989). Test acustico-percettivo su voci di gemelli. In *Proceedings of the xvii convegno nazionale AIA*, Parma, Italy, 12-14 april 1989, 427-432.

Giannini, A., Pettorino M. (1992) *La fonetica sperimentale*. Napoli: Liguori, 197.

Gedda, L., Bianchi, A. & Bianchi-Neroni, L. (1958). La Voce dei Gemelli—I. Prova di identificazione intrageminale della voce in 104 coppie (58 MZ e 46 DZ). In *Acta geneticae medicae et gemellologiae: twin research*, 4(2), 121-130.

Gonzalez, J., Cervera, T. & Llau, M.J. (2003). Acoustic analysis of pathological voices compressed with MPEG system. In *Journal of voice*, 17(2), 126-139.

Heeringa, W., Van de Velde, H. (2017). Visible Vowels: A Tool for the Visualization of Vowel Variation. In International Speech Communication Association (ISCA) (Ed.) (2017). *Proceedings of INTERSPEECH 2017*, Stockholm, August 20–24, 4034-4035.

Jensen, K. (2002). The timbre model. In *Journal of the Acoustical Society of America*, 112(5), 2238-2238.

Johnson, K., Azara, M. (2000). The perception of personal identity in speech: evidence from the perception of twins' speech. Unpublished manuscript.

Kader, F., Ghai, M. & Olaniran, A.O. (2020). Characterization of DNA methylation-based markers for human body fluid identification in forensics: a critical review. In *International journal of legal medicine*, 134(1), 1-20.

Künzel, H.J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. In *Forensic Linguistics*, 8(1): 80–99.

Künzel, H.J. (2002) 'Rejoinder to Francis Nolan's "The 'telephone effect' on formants: a response"', *Forensic Linguistics*, 9(2): 83–6.

Künzel, H.J. (2010). Automatic speaker recognition of identical twins. In *International Journal of Speech, Language, and the Law*, 17(2), 251-277.

Loakes, D. (2008). A forensic phonetic investigation into the speech patterns of identical and non-identical twins. In *International journal of speech, language, and the law*, 15(1), 97-100.

Lobanov, B.M. (1971). Classification of Russian Vowels Spoken by Different Speakers. In *Journal of the Acoustical Society of America*, 49(2), 606-608.

Nolan, F., Oh, T. (1996). Identical twins, different voices. In *International Journal of Speech, Language, and the Law*, 3(1), 39-49.

Planterose Jiménez, B., Liu, F., Caliebe, A., Montiel González, D., Bell, J.T., Kayser, M. & Vidaki, A. (2021). Equivalent DNA methylation variation between monozygotic co-twins and unrelated individuals reveals universal epigenetic inter-individual dissimilarity. In *Genome Biology*, 22(18): doi:10.1186/s13059-020-02223-9

Prieto, P., Borràs-Comes, J. & Roseano, P. (coords.) (2010-2014). Interactive atlas of romance intonation, retrieved from http://prosodia.upf.edu/iari/.

Rietveld, A.C.M., van Heuven, V.J. (2009). *Algemene fonetiek*. Bussum: Uitgeverij Coutinho.

Rhodes, R.W. (2012). Assessing the strength of non-contemporaneous forensic speech evidence, PhD dissertation, University of York.

San Segundo, E., Künzel, H. (2015). Automatic speaker recognition of Spanish siblings: (monozygotic and dizygotic) twins and non-twin brothers. In *Loquens*, 2(2), e021-e021.

San Segundo, E., Tsanas, A. & Gómez-Vilda, P. (2017). Euclidean distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics. In *Forensic Science International*, 270, 25-38.

Sebastian, S. (2013). An investigation into the voice of identical twins. In *Otolaryngology online journal*, 3(2), 9-15.

Van Braak, P., Heeren, W.F.L. (2015). "Who's calling, please?" Is there speaker-specific information in twins' vowels?. In *24th Annual Conference of the International Association for Forensic Phonetics and Acoustics*, Leiden, Netherlands, 7-10 July 2015, conference presentation.

Van, W.G., Vercammen, j. & Debruyne, F. (2001). Voice similarity in identical twins. In *Acta oto-rhino-laryngologica belgica*, 55(1), 49-55.

Vidaki, A., López, C.D., Carnero-Montoro, E., Ralf, A., Ward, K., Spector, T. & Kayser, M. (2017). Epigenetic discrimination of identical twins from blood under the forensic scenario. In *Forensic Science International: Genetics*, 31, 67-80.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. In *The Journal of the Acoustical Society of America*, 69(5), 1465-1475.

Zuo, D., Mok, P.P.K. (2015). Formant dynamics of bilingual identical twins. In *Journal of phonetics*, 52, 1-12.