DOI: 10.17469/O2107AISV000019

SONIA CENCESCHI, CHIARA MELUZZI, NICHOLAS NESE

Speaker's identification across recording modalities: a preliminary phonetic experiment[1]

This work investigates how WhatsApp audio messages could be compared to high quality professional recordings and low quality ones in a forensic framework. A controlled experiment with 12 Italian students (6F, 6M) was performed in order to ascertain whether formants' values of the three cardinal vowels /a/-/i/-/u/ will help in distinguishing the same speaker across three different recording modalities. Both unnormalized data in Hertz and normalized values (Lobanov and Bark) were compared across the male and female subsets. Results indicate that unnormalized data performed better than normalized ones, and that a qualitative investigation has to be combined with a quantitative one. This preliminary work opens the way to further investigations on the possibilities of WhatsApp audio messages for forensic purposes at the crossroads between linguistics and engineering.

Keywords: forensic phonetics, WhatsApp, forensics linguistics, environmental recordings, vowels' formants, intra-speaker variation.

1. *Introduction*

This work addresses a central issue in forensic phonetics: the possibilities of identifying a speaker's voice through different recording modalities. With the growth of environmental recordings (Orletti & Mariottini, 2017; Ministero della Giustizia, 2018), and the spread of WhatsApp technology a new challenge opens for both linguists and engineers working in the forensic field. To put it simply, how is it possible to ascertain the identity of a speaker across different recordings, especially when those recordings are highly deteriorated? The growing interest in the possibilities of automatic or semi-automatic comparisons have usually worked on a suprasegmental or acoustic level, by leaving aside linguistic variation as expressed (Drygajlo, Jessen, Gfroer, Wagner, Vermeulen & Niemi, 2016; Tirumala, Shahamiri, Garhwal & Wang, 2017; Jagdale, Shinde & Chitode, 2020). At the same time, sociophonetics researches have demonstrated the great heuristic power of vowels' formants in explaining variabilities across and within speakers (e.g., Quené, 2008).

[1] This work has been conceived and written jointly by the three authors. However, for the Italian evaluation system, author 1 is responsible for section 2, 3, and 4.2 (with subsections); author 2 is responsible for sections 1, 4, 4.1, 5 and 6; author 3 is responsible for sections 4.3 and 7. Authors 1 and 2 have also conceived the experimental design of the work and analyzed the data, whereas author 3 performed the data collection and the annotation of the whole corpus.

In this respect, this work proposes a preliminary phonetic experiment focused on vowels' variation as produced by 12 speakers (6F, 6M) recorded in three different settings: a professional high quality recording, a WhatsApp audio message, and an environmental low quality recording.

The paper is organized as follows: section 2 presents the theoretical premises of the work, and section 3 the research questions we aimed at answering in this paper. The experimental design, the research protocol, and the corpus with the associated technical issues will be presented in section 4 (and subsections).

The fifth section presents the results of both a qualitative (5.1) and a quantitative (5.2) analysis on F1 and F2 of the three cardinal vowels /a/-/i/-/u/, with a summary of results (5.3) for the individual variation. Finally, section 6 discusses the results in light of their possible applications in forensic phonetics, and section 7 presents our first conclusions and further perspectives.

2. *Some theoretical premises*

Audio forensics daily deals with spontaneous speech, but the specialized research suffers from a lack of spontaneous speech corpora, mainly due to privacy reasons. As already introduced in Cenceschi, Trivilini, Sbattella & Tedesco (2019), we think that social media apps could represent a very large digital pool from which to draw (Kaplan, 2015). They constitute a fundamental part of modern human communication, increasing enormously in recent years among users of all ages. However, for the purposes of both linguistic and forensic analysis, it is necessary to investigate similarities and variations with respect to phone-calls and live speech, because the audio message as a category belongs to a new speech communication style (Nencioni, 1983; Cenceschi, Sbattella & Tedesco, 2018). As a consequence, audio messages have introduced new interaction behaviors such as different speakers' expectations, rhythm, pauses, etc. From a technical point of view, audio messages also provide data with different qualities, if compared to laboratory recordings.

Through the years, a large variety of spontaneous Italian speech corpora have been collected by scholars (Cresti, Moneglia, do Nascimento, Moreno-Sandoval, Véronis, Martin & Blum, 2002; Albano Leoni, 2006; Cresti & Panunzi, 2013). However, none of them focused on social media speech style. The inter-device speech features variability has also been investigated by various works, in order to highlight any differences between the audio in the various compressed formats (Nolan, Grigoras, 2005; Khan, Wiil & Memon, 2010; Gold, French, & Harrison, 2013; van Braak & Heeren, 2015). It remains up to investigate whether this two dimensions co-occur in shaping the variability of speech data, that is to say how speech varies both across style (and, in particular, social-media styles) and recording modality. In this respect, some explorative studies have addressed the issue of intra-speaker and inter-devices variations for a limited number of speakers. For instance, Cenceschi et al. (2018) consider only 2 speakers, and compare WhatsApp

data with phone-calls, in order to verify the hypothesis that these data will be equivalent for most of the parameters despite the diatechnical variations.

The forensic consequence will be that WhatsApp data can be used as a parallel or even alternative source in forensic investigations.

However, the main difference between recording modality was of prosodic and temporal nature, thus leaving an open hypothesis on what happens on the segmental level.

3. *Research questions and main objectives*

Given the lack of studies considering speech variation between recording modalities from a linguistic perspective, we decided to start with a first explorative study to investigate the intra-speaker and inter-device variations for a limited number of speakers. The main purpose of this paper is to understand whether and to what extent is possible to compare WhatsApp data with high quality and low quality recordings. In particular we will consider if different audio typologies affect the main phonetic characteristics for voice identification in forensic settings. Therefore, our specific research questions for this first study are:

1. Are recordings made with different devices directly comparable among them?
2. What is the impact of common normalization procedures on comparability between different recording formats?
3. To what extent is, then, possible to compare/identify the voice of the same speaker from different recording devices in forensic analysis?

Indeed, when working in forensics it frequently happens to compare two audio samples recorded in different modalities. The usual request from the law forces to the expert (i.e., the phonetician) is to ascertain whether the two samples could belong to the same speaker.

Conversely, in the present experiment, we are certain that the voices belong to the same speakers recorded in different modalities. The main aim of our work, thus, is to verify if a semi-automatic investigation of some phonetic features will confirm that the samples belong to the same voice. If the results will lead to the emergence of a difference in speakers' identification according to the recording devices, this will tell something important for what it concerns the possibilities of vocal comparison in phonetic forensics.

4. *The WAsp Corpus*

Our preliminary study consisted in the creation of a small corpus (called *Wasp*) based on the productions of 12 speakers (6M, 6F) with the same sociolinguistic characteristics. They were all students enrolled at the University of Pavia in various courses, with a preference for non-linguists. All students were born and living in the north-west of Italy, and they were Italian L1; no bilingual students have been included, although all participants have knowledge of various foreign languages

(e.g., English). All speakers were consciously and freely taken part in the experiment as volunteers, receiving no compensation for their participation in the project. Speakers personal data were anonymized and protected according to the current ethical and privacy dispositions. Ethical and privacy agreement was signed by both the researchers and the participants.

Each speaker in the Wasp corpus performed two different tasks in three different recording conditions. The two tasks consist in a sentence-list reading of 30 sentences, with a pause of about 3" between each sentence, and a description task, thus producing short monologues about the furniture of their room and the explanation of the cooking of their favorite food.

Each speaker has asked record with three different modalities (audio formats are deepened in the related paragraph):

1. Recorded by an expert, simulating a high quality comparative forensic registration in a sound-proof environment, by means of a Tascam DR-05.
2. Auto-recording of a WhatsApp audio message.
3. Through a phone call made by the researcher and recorded through the App Voice Recorder 2.81 in *.mp3*.

During each recording session, the speaker was asked to repeat the sentence-list reading and the description tasks twice, with a short pause between the two recordings, in partial accordance with the Protocol for the collection of databases of forensic recordings (Morrison, Rose & Zhang, 2012).

Speakers were recorded by the third author in a soundproof room at the boarding school 'Giasone del Maino' in Pavia, in the afternoon of 4th December 2019. The boarding school was chosen because all participants were hosted there, and the soundproof room offered the ideal environment for high quality recordings.

After this first session, speakers were asked to reach their private room and, when they felt comfortable, send a WhatsApp message to the second author by following the same protocol (i.e., reading list and short description, each one repeated twice, possibly in two different messages). When receiving the messages, the researcher controlled them for their completion (e.g., presence of both the tasks, and of the two repetitions), eventually soliciting the repetition of a task.

Finally, the second researcher called each participant and recorded them repeating the two tasks twice though the App Voice Recorder installed on her phone. The speakers were asked to receive the phone calls in a possibly silent environment like their personal bedrooms, and the researcher made the calls from her office, which was not a soundproof room.

The data were, thus, acquired in three different moments, albeit very close to each other. It was judged difficult to record the same speaker at a single moment in the three modalities, because the mobiles would have created interferences with the microphone. We do not believe that speakers will change much in style while performing a reading task, and this was also one reason for opting for a reading task vs. a real dialogical task: as a preliminary research, our main interest was only devoted to recording modalities, without other variables involved. The data acquired

according to these three different modalities were stored according to task, and by identifying each speaker through an alphanumeric label in order to later compare their speech through the different recording modality.

4.1 Task and acoustic features

The sentence-reading task contains 12 target words, balanced by target stressed vowels /a/, /i/ and /u/ followed by a singleton or geminate alveolar or bilabial consonant (e.g., *Papa*, *pappa*, *Tita*, *Titti*). An equal number of fillers was also added in each list. The list was presented to the speaker in a randomized order.

Target words and fillers appear in sentences with similar prosodic contour (e.g., *La tata guarda i bimbi al parco* "The nanny looks after the kids in the park").

The similarity of prosodic profile was aimed at reducing the involved variables as much as possible, to ensure repeatability, and to lay the ground for future studies gradually introducing further parameters (e.g., enlarge the dataset to sentences containing a pragmatic accent to understand how its characterization changes according to the recording methodology). According to the proposal of Cenceschi *et al*. (2018) and van Braak & Heeren (2015), we limited the study to read speech and short spontaneous descriptions, without pragmatic accents, emotions, dialog interaction, and by maintaining the same talking speed (as far as possible) in the three settings.

For the same reasons, the recording must be realized in a silent room, but without specific details regarding the environmental soundproofing. The aim is to simulate a recording in normal everyday life, without any particular noisy conditions that will be introduced in future studies.

4.2 Technical equipment and digital formats

The three recording modalities produce different audio outputs whose digital quality mainly depend on: hardware equipment (Microphone, CPU, chipsets, etc.), and compression formats. The different characteristics are shown below with those of the professional microphone.

4.2.1 Hardware smartphone equipment

The 12 speakers have phones of different brands, and specifically: 5 iPhone (3 iPhone-7, 2 iPhone-8, 1 iPhone-11), 2 Huawei (1 P10 and 1 P20), and one unit for Xiaomi (Redme note7), Asus (Z00ED), Google (pixel 2 XL), Honor (version 10), and Samsung (J330FN). Although the technology of the different smartphone models can cause quality variations, albeit minimal to the human ear, this variable will not be investigated here as the sample is too limited. Then, we consider the possible discrepancies momentarily irrelevant in the context of a macroscopic analysis. A future enlargement of the sample will allow further analyzes concerning the variability between the various categories.

4.2.2 Professional microphone

The high quality recordings were performed with a Tascam DR-05, without any external microphones connected to the recorder. Its sound quality guarantees over 92dB signal to noise ratio, under 0.05% total harmonic distortion and 20Hz to 40kHz response (-1/+3dB) at 96kHz/24-bit resolution. Its output has been setted as *wav* at 44.100 Hz - 16 bit in order to simulate the higher quality recordings typical of forensic investigations.

4.2.3 Format compression and conversion

The difference between low and high quality file formats depends on the modality used to encoding or decoding a digital data stream. The corpus comprises four different data formats: *wav*, *mp3*, *ogg*, and *m4a* depending on the recording modality:

- Professional recordings: uncompressed *wav* files 44.100 Hz - 16 bit.
- Voice Recorder 2.81 *mp3*, 16.000 Hz - 128 kbps, codec Lame..
- WhatsApp: *ogg*, 64 kbps, codec Vorbis.
- WhatsApp: *m4a* 64 kbps, codec AAC.

WhatsApp is based on the SILK VoIP (Voice over Internet Protocol) codec developed by Skype and now licensed out, being available as open-source freeware: the voice (and other media) are delivered over an IP packet switched network. It is a foundation, with CELT, of the hybrid codec Opus. WhatsApp exports now Opus files with pseudo file extension *m4a* (AAC codec) or as *ogg* (Vorbis codec) because from about 2018, the Opus format is not recognized by many apps. The VoIP technique has a major impact on the spectro-acoustic properties of the signal (as already addressed for automatic speaker recognition in Khan, Baig & Youssef (2010) because it was introduced to preserve the network bandwidth to the detriment of signal quality (Singh & Mian, 2016). Voice Recorder version 2.99 by Splend App works for Android 4.1+. It allows a variable bitrate from 32 up to 320 kbps, and sampling rate from phone quality (8 kHz) to CD quality (44 kHz). It has been setted with *mp3* 16 kHz - 128 kbps. The app has been installed on one author's smartphone and started just before the phone calls with the speaker. It exploits the microphone of the mobile phone on which it is installed: as a consequence the two voices have unbalanced intensity, and the quality of the speaker's speech is very limited with respect to other recording modalities. This phenomenon well approximates the worst forensic recordings, as realized in real context scenarios. Furthermore, it should be remembered that private dialogues are often recorded with similar modalities, without adequate technological knowledge, and through applications found on the web. The *ogg* and *m4a* recordings have been converted to *wav* in order to allow the Praat analysis as usually performed in linguistics (De Decker, Nycz, 2011; Styler, 2013). As underlined in (Wang et al. 2018), the up-sampled recording does not modify speech features because acting a PCM linear conversion form a compressed format to a higher quality one.

4.3 The corpus

We base our present analysis on 864, that is 12 target words, twice repeated in three different recording settings by 12 different speakers. The corpus is balanced for sex of the speaker, recording condition, target vowels (/a/, /i/, /u/), phonotactic environment (singleton, geminate), and surrounding consonants (either bilabials or alveodentals, in both cases voiceless). Each target word was manually annotated on three different tiers in PRAAT: a first tier included the whole sentence, whereas on the second tier the target word was isolated, and on the third tier we segmented the target vowel and the following consonant.

For setting the vowel's left and right boundaries we base on the beginning and end of the second formant (F2); the occlusive consonants include the whole silence phase and the following VOT, and it ends when the F2 of the following vowel appears.

After the annotation, we automatically extracted the following acoustic parameters: pitch, F0, F1 and F2 of the target vowels, jitter and shimmer, duration of both the target vowel and the following consonant. It should be mentioned that we extracted the formants' values twice: at the midpoint, at five different timepoints through the whole segment. Although in this paper we will focus only on formants' variations in a static approach (i.e., by looking at midpoint values), further research will include a dynamic approach on formants' variation across the segment, together with jitter and shimmer analysis. The data have been inserted in a matrix on the software IBM SPSS 20, and also visually inspected through the web application Visible Vowels (Heeringa & Van de Velde, 2018).

5. *Analysis*

Since the preliminary nature of this work, the dataset was balanced but limited in the amount of samples, so that a detailed statistical analysis with all the possible variables (e.g., phone label) can't be performed. The analysis will thus consist of a first qualitative analysis, performed through the inspection of formants' variations and vowel space variation through Visible Vowles, and a second quantitative analysis on formants' variation by performing different Anovas on IBM SPSS 20.

5.1 Qualitative analysis

A first visual inspection of formants' variation in the three recording modalities highlights some differences with respect also to vowel quality. In the graphs, as well as in the following statistical analysis, we maintain as separated the values for males and females in our corpus, because of the notorious biological differences affecting formant values (especially in non-normalized data). The three different recording modalities will be indicated as MIC for the high quality recordings made by the expert, WA for the WhatsApp audio messages sent by the participants, and VR for the phone call recordings made through the Voice Recorder Application.

Figure 1 - *Graphic representations of mean values variation of F1 (above) and F2 (below), with unnormalized data, in the three recording modalities (MIC = microphone, VR = Voice Recorder, WA = WhatsApp) divided by speakers' sex. The visualization method indicated as TL is based on Fox & Jacewizc (2009)*

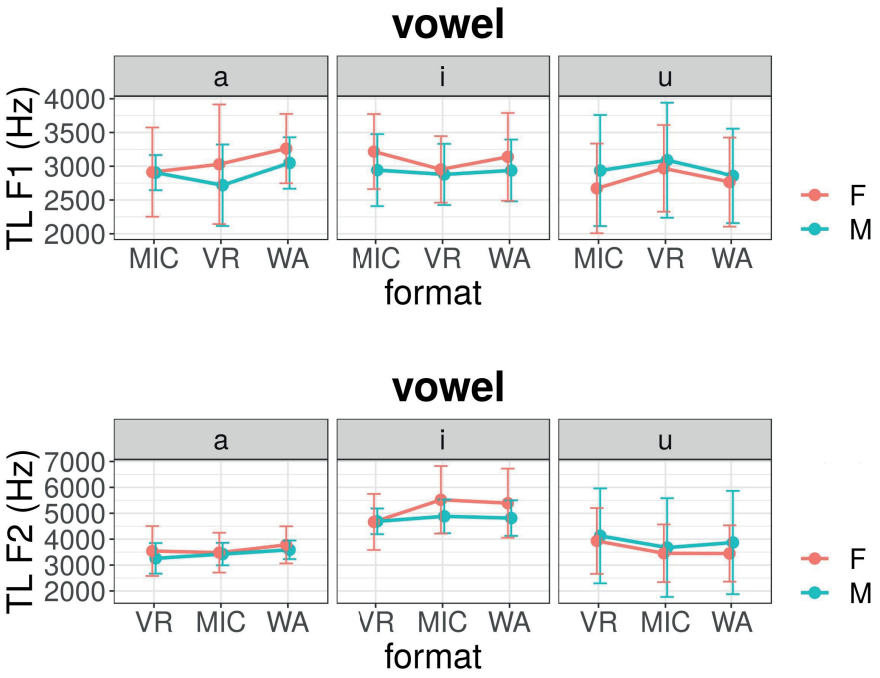
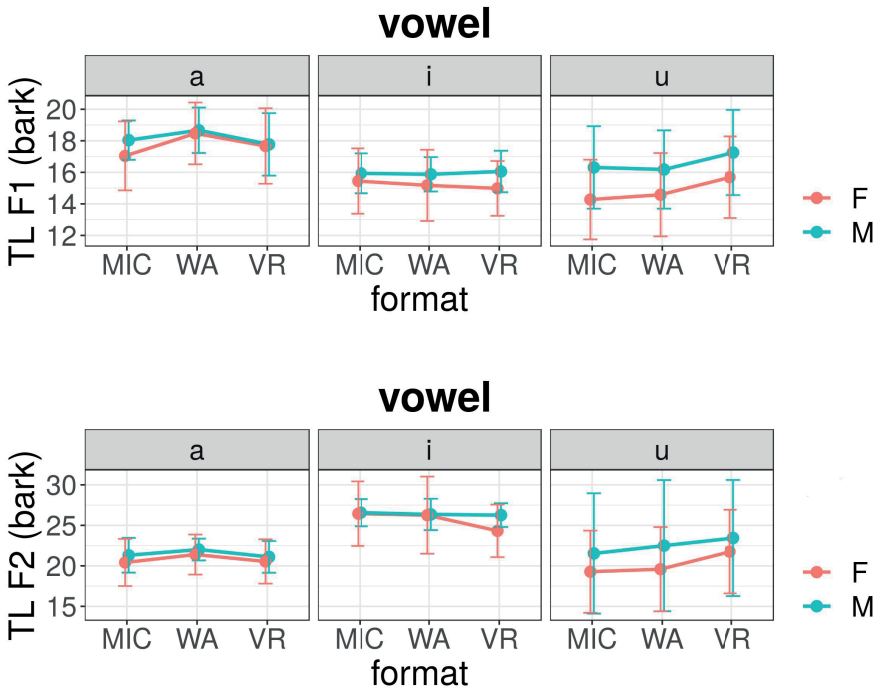


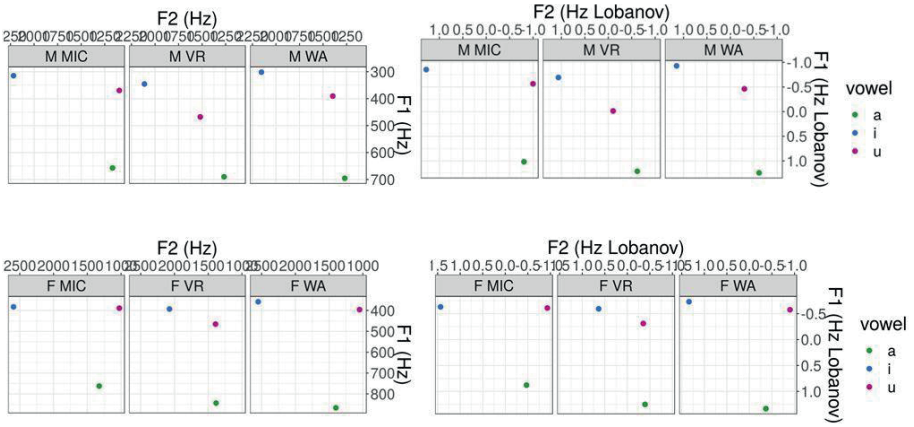
Figure 1 presents the variation of the mean values in the female and male subgroups according to the three recording modalities. It appears that the variation is, in general, quite minimal, with a similar behavior between males and females, but with some differences with respect to the formant and the vowel quality. For instance, as it concerns F1, it is possible to notice how values are higher for /u/ in the low quality recordings (VR) than in the two other modalities. Conversely, for /i/, F1 values are lower in the VR modality than in either high quality recordings and WhatsApp message. As for F2, an overall similarity is noticeable for /a/, whereas for the two high values there is again a difference between the low quality recordings (VR), on the one side, and the high quality recordings and WhatsApp messages, on the other side. Thus, from this preliminary investigation it appears that the low quality recordings as performed through phone call registration present the most different values, whereas the mean values emerging from WhatsApp audio messages are similar to the ones recorded in a professional environment.

Figure 2 - *Graphic representations of mean values variation of F1 (above) and F2 (below), with data normalized in Bark (Traunmüller, 1990), in the three recording modalities (MIC = microphone, VR = Voice Recorder, WA = WhatsApp) divided by speakers' sex. The visualization method indicated as TL is based on Fox & Jacewizc (2009)*



In Fig. 2 the same data are presented, but the values of F1 and F2 have been normalized in Bark through Traunmüller's (1990) formula. As it appears from the graphs above, for both formants the variation is minimal for the vowel /a/, although there is a greater dispersion in F2 for male subjects. As noticed for non-normalized data, both high quality recordings (MIC) and WhatsApp messages (WA) show similar values if compared to low quality recordings (VR).

Figure 3 - *Vowel space representation of variation of the three cardinal vowels /a/-/i/-/u/ in the three recording modality (MIC = microphone, VR = Voice Recorder, WA = WhatsApp), with respect to speaker's sex (M = males, above; F = females, below). Unnormalized data are presented on the left, normalized data with Lobanov's (1971) formula are presented on the right*



We also compare the position of the three cardinal vowels accordingly to their mean values of F1 and F2, with data unnormalized and normalized through Lobanov's (1971) formula (Fig. 3). Indeed, many studies in sociophonetics have highlighted how Lobanov's (1971) normalization procedure should be preferred because it preserves more information on the socio-indexical dimension (cf. Van der Harst, 2011; Adank, 2003). However, this normalization procedure usually works better with huge corpora: this is not the case either in our work nor, more generally, in forensics.

Data presented in Figure 3 show little or no difference between unnormalized mean values and normalized ones, both in the male subgroup (above) and in the female one (below). Moreover, once again it seems possible to highlight a major similarity between high quality recordings (MIC) and WhatsApp audio messages (WA). However, the values for /a/ represent an exception since WA modality shows a lower F2 and a higher F1 than in both MIC and VR. Conversely, VR setting seems to play a major role on the posterior vowel /u/, which appears to be more centralized than in MIC and WA settings. In the female subgroup, there seems also to be an influence on the values of /i/, which appears to be more centralized in VR recordings than in the other two settings.

To sum up, this first qualitative inspection of our data has pointed out some important points. Firstly, the variation between recording modalities doesn't appear to be a huge one, in particular between professional recordings and WhatsApp audio messages. In this respect, the phone call recordings realized with the app Voice Recorder are more dissimilar from the other two settings, in particular for what it concerns the two extreme vowels /i/ and /u/. Furthermore, it has been observed that Lobanov's normalization does not differ much from unnormalized data; for this reason, it has been decided to exclude this procedure from the next quantitative analysis.

5.2 Quantitative analysis

After a qualitative analysis of the different formant values across the three recording modalities, the question arises on whether these differences are only descriptive or not. In other words, we would like to investigate if a statistical analysis on formants' values would certify that the speaker is the same across the different recording modalities. Indeed, if the test will result significantly (i.e., $p < 0.05$), it will mean that the difference between the two (or three) settings is so huge to be attributed to different speakers. Conversely, if no significance will be found, it will mean that, albeit some differences due to the quality of the recordings, it is still possible to recognize that the different vowels belong to the same speakers.

We run Anovas on both unnormalized values and values normalized in Bark. Since male and female formant values are very different due to biological reasons, in both cases we maintain the two subgroups separated. A post hoc Tukey test was also performed on both F1 and F2 values, in Hertz and in Bark, in order to verify if differences between the recording modalities are statistically significant.

From the data it emerges that in both male and female subgroups the comparison of formant values in the three recording modalities is always statistically significant ($p < 0.05$). An exception is represented by the F2 of the vowel /a/, but only in the female subgroup ($p = 0.063$).

By looking at the post hoc Tukey test results, the following picture emerges for the two subgroups. For the female speakers, F1 unnormalized values are not significant for /a/ between WhatsApp and Voice Recorder settings, for /i/between high quality recordings and the other two modalities, and for /u/ only between high quality recordings and WhatsApp messages. In this subgroup, F2 values are significantly different among all devices for /a/, but only in comparing high quality recordings and WhatsApp messages for /i/ and /u/. Conversely, no significance has been found in the cross-modalities comparison with Bark values, with the exception of the MIC-WA comparison for the F2 of /i/ and /u/.

For the male group, F1 values for /a/ were statistically significant in comparing WhatsApp messages and Voice Recorder, both with unnormalized and normalized Bark data. However, for /i/ and /u/ only unnormalized data show a difference between high quality recordings and WhatsApp audio messages. The same could be said for the F2 of /i/ and /u/, both for unnormalized and normalized Bark data. With Hertz value a difference also has emerged between WhatsApp and the Voice Recorder App for the F2 of /i/. Finally, with Bark data, there was always statistical significance in comparing the three recording modalities when considering the F1 of both /i/ and /u/.

As for individual variation, unnormalized data predicted better the coincidence of the speaker across recording modalities, whereas normalized Bark data always resulted in a statistical significant difference, with but 1 exception (cf. Appendix). Indeed, the ANOVA also shown that for some speakers unnormalized data could predict their identity across the three recording modalities (e.g., speaker CM), and

that the back vowel /u/ preserved better this individual difference, with respect to both /i/ and /a/.

To sum up, it appears that normalized data performed worse than unnormalized data in recognizing the same voices across recording modalities. Among vowels, /a/ seems to create major confusion, especially for what it concerns F1 values. Conversely, the vowel /i/ appears to better perform in recognizing the three recording modalities as belonging to the same speakers. Moreover, the two low quality recording modalities (i.e., WhatsApp and Voice Recorder) generate more confusion than the comparison between a low quality recording (especially WA) and high quality one (MIC).

6. *Discussion*

This work has focused on a common problem in forensic phonetics: the comparability and recognizability of speakers across different speech samples, recorded with different modalities. This problem has usually been addressed from an engineering point of view (cf. 4.2.3): for instance, Khan et al. (2010) and Singh et al. (2016) proposed a semi-automatic speech recognition system by considering the loss of spectral information. In this work, we intended to address the issue from a linguistic point of view, with a phonetic analysis of vowels' formants variation across recording modalities.

The results of both our qualitative and quantitative analysis both point at a major reliability of direct values, without normalization. As expected, the low quality recordings realized with the app Voice Recorder badly performed because of the high formants variability. It is important to stress that this represents the typical forensic case, when an environmental interception (similar, for quality, to our VR setting) has to be compared with professional recordings (our MIC setting). Therefore, our results suggest that this comparison should be addressed with a semi-automatic analysis only with extreme caution. A combination of quality and quantity analysis seems to be preferable, especially when working with a small dataset, as it frequently happens in phonetic forensics. For what it concerns the different vowels, the central vowel /a/ seems to be more indicated for comparing low-quality audio files (such as VR and WA). Conversely, extreme vowels /i/ and /u/ seem more suitable for forensic comparison, especially between low quality recordings and high quality ones. However, this could be a language-specific difference that should be tested on inter-linguistically with a similar research protocol.

Finally, it has been repeatedly noted how normalization prevents recognition of different samples as belonging to the same voices across different recording modalities. Bark normalization presents some exceptions, but with a considerable variability not only among the cardinal vowels considered but also between the male and female subgroups. The Lobanov normalization was completely ineffective or counter-productive in our case, but this is probably due to the small dimension of our corpus, since Lobanov's formula squeezes the values too much. However, this

lack of informativity of normalized data could also depend on the target approach chosen for this analysis, with values extracted on the midpoint of the stressed vowels. Although this is the most common practice in forensic phonetics, it is true that (socio)phonetic analysis nowadays relies more on dynamic approaches for vowel analysis (e.g., Farrington et al. 2018, van der Harst et al. 2014).

7. *Conclusions and further perspectives*

Forensic linguists are always asked to ascertain the identity of a speaker across short speech samples frequently recorded with different sound qualities. In this paper, we proposed a laboratory experiment aimed at exploring the possibilities of (semi) automatic comparison of formant values. We recorded 12 speakers in three different modalities (i.e., professional high quality audio, WhatsApp messages, low quality environmental recordings). The values of F1 and F2 of the three cardinal vowels /a/-/i/-/u/ were compared both qualitatively and quantitatively, and with and without normalizing the original Hertz data. The analysis allows us to answer our research questions, by also opening the fields for further discussions and experiments on this topic.

Indeed, we show that audio files made with different recording devices, and especially low and high quality ones, suggest to combine a qualitative inspection of data distribution through graphic representations (e.g., with the online software Visible Vowels) with a statistical comparison of formants' values. Normalization procedures usually adopted in (socio)phonetic analysis (i.e., Bark and Lobanov) do not work well with small subsets like the ones commonly available for forensic comparisons. In particular, Lobanov squeezes the values too much for allowing a comparison, whereas Bark normalized data are randomly significant in recognizing the same speakers across recording modalities. In particular when working with extremely compromised audios, a substantial precautions in linguistic and phonetic analysis is needed. Comparing the voices of a possible same speaker from different recording devices for forensic purposes is possible, but a qualitative analysis has to be combined with a quantitative one.

Furthermore, WhatsApp audio messages turn out to be a good compromise between good quality (professional) recordings and low quality (environmental) ones. This leads to hypothesize that their use in forensic phonetics will increase in the future, also because of their availability. From a linguistic point of view, WhatsApp messages could also be said to represent a new form of expressive modality (Nencioni, 1983), conceivable halfway between spontaneous and recited speech.

Obviously, further experiments should be conducted on more spontaneous samples, since a word list reading task is quite different from real speech from a stylistic and a phonological point of view, as it has been pointed out by previous scholars (cf. 2). However, a preliminary investigation like the one proposed here was necessary to ascertain the difference between recording modalities without the 'noise' generated by variability in spontaneous speech. Further studies will, thus, address

other phonetic and phonological variables (e.g., prosodic contour), and widen the analysis to spontaneous social-media speech. It will also be desirable to confirm production analysis with perceptive tests, in order to verify whether and to what extent recording modalities affect our capability to recognize the speakers, especially with extremely deteriorated recordings.

All these issues are extremely important for the practical application of linguistic analysis to forensics. Although until now they have been scarcely addressed, as far as we know, from a linguistic point of view, much work has been done by engineers. Therefore, an interdisciplinary approach will benefit the investigation and strengthen the results on speakers' semi-automatic comparability.

Bibliography

Adank, P.M. (2003). *Vowel Normalization. A Perceptual acoustic study of Dutch Vowels*. Netherlands Graduate School of Linguistics: LOT.

Albano Leoni, F. (2006). *Il corpus CLIPS*, presentazione del progetto. Dostupno na: http://www. clips. unina. it/it/[25.10. 2012].

Cenceschi, S., Sbattella, L. & Tedesco, R. (2018). Verso il riconoscimento automatico della prosodia. In *STUDI AISV*, 433-440.

Cenceschi S., Trivilini A., Sbattella L. & Tedesco R. (2019) *Collecting Italian spontaneous social media speech: the WAsp2 project*, AISV conference 2019, Arezzo (Italy).

Cresti, E., Moneglia, M., do Nascimento, F.B., Moreno-Sandoval, A., Véronis, J., Martin, P. & Blum, C. (2002). The C-ORAL-ROM Project. *New methods for spoken language archives in a multilingual romance corpus*. In LREC.

Cresti, E., Panunzi A. (2013). *Introduzione ai corpora italiani*. Bologna: Il Mulino.

De Decker, P., Nycz, J. (2011). For the record: Which digital media can be used for sociophonetic analysis?. *University of Pennsylvania Working Papers in Linguistics*, *17*(2), 7.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2016). *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*. Verlag für Polizeiwissenschaft.

Farrington, C., Kendall, T., & Fridland, V. (2018). Vowel Dynamics in the Southern Vowel Shift. In *American Speech: A Quarterly of Linguistic Usage*, 93(2), 186-222.

Fox, R.A. & Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. In *The Journal of the Acoustical Society of America*, 126(5): 2603-18.

Gold, E., French, P. & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. In *Proceedings of Meetings on Acoustics*, 19, 1-7.

Heeringa, W., Van de Velde, H. (2018). Visible Vowels: a Tool for the Visualization of Vowel Variation. In *Proceedings CLARIN Annual Conference 2018, 8 - 10 October, Pisa, Italy*. CLARIN ERIC.

Jagdale, S.M., Shinde, A.A., & Chitode, J.S. (2020). Robust Speaker Recognition Based on Low-Level-and Prosodic-Level-Features. In J. Vanita, G. Chaudhary, M.C.

Taplamacioglu & M.S. Agarwal (eds.) *Advances in Data Sciences, Security and Applications*, Singapore: Springer, pp. 267-274.

Kaplan, A.M. (2015). Social Media, the Digital Revolution, and the Business of Media. In *International Journal on Media Management*, 17(4), 197-199.

Khan, A., Wiil, U.K., & Memon, N. (2010). Digital forensics and crime investigation: Legal issues in prosecution at national level. In *2010 Fifth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering* (pp. 133-140). IEEE.

Khan, L.A., Baig, M.S., & Youssef, A.M. (2010). Speaker recognition from encrypted VoIP communications. In *Digital investigation*, *7*(1-2), 65-73.

Ministero della Giustizia (2018). *Relazione del Ministero sull'amministrazione della giustizia*, p. 25.

Morrison, G.S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic- voice-comparison research and practice. In *Australian Journal of Forensic Sciences*, 44(2), 155- 167.

Nencioni, G. (1983). *Di scritto e di parlato*. Bologna: Zanichelli.

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. In *Journal of Speech, Language and the Law*, 12, 143-173.

Orletti F. & Mariottini L. (2017). *Forensic Communication in Theory and Practice: A Study of Discourse Analysis and Transcription*, Cambridge: Cambridge Scholars Publishing.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. In *The Journal of the Acoustical Society of America*, *123*(2), 1104-1113.

Singh, H., Mian, M. (2016). Comparative Study and Analysis of various VoIP coding Algorithms. In *International Journal of Computer Applications*, 141(2).

Styler, W. (2013). Using Praat for linguistic research. *University of Colorado at Boulder Phonetics Lab*.

Tirumala, S.S., Shahamiri, S.R., Garhwal, A.S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. In *Expert Systems with Applications*, 90, 250-271.

Traunmüller, H. (1990) Analytical expressions for the tonotopic sensory scale. In *Journal of the Acoustical Society of America,* 88 (1), 97-100.

van Braak, P., Heeren, W.F.L. (2015). *"Who's calling, please?" Is there speaker-specific information in twins' vowels?*, Bachelor's thesis, Utrecht University.

van der Harst, S. (2011). *The vowel space paradox: A sociophonetic study on Dutch*. Netherlands Graduate School of Linguistics: LOT.

van der Harst, S., Van de Velde, H., & Van Hout, R. (2014). Variation in Standard Dutch vowels: The impact of formant measurement methods on identifying the speaker's regional origin. In *Language Variation and Change*, *26*(2), 247-272.

Wang, Z., Yan, D., Wang, R., Xiang, L., & Wu, T. (2018). Speech resampling detection based on inconsistency of band energy. *CMC-Comput., Mater. Continua*, *56*(2), 247-259.

Appendix

Anova on invididual variation across recording modalities

Speaker	Vowel	Formant	Normalizatin	Anova F(2,21)	p value
AB_F	a	F1	None (Hertz)	45.855	0.0001
			Bark	80.179	0.0001
		F2	None (Hertz)	2.217	**0.134***
			Bark	86.47	0.0001
AM_F	a	F1	None (Hertz)	24.774	0.0001
			Bark	36.219	0.0001
		F2	None (Hertz)	16.997	0.0001
			Bark	1.194	**0.323***
CM_F	a	F1	None (Hertz)	30.793	0.0001
			Bark	137.379	0.0001
		F2	None (Hertz)	5.179	0.15
			Bark	132.519	0.0001
DS_M	a	F1	None (Hertz)	6.712	0.06
			Bark	77.347	0.0001
		F2	None (Hertz)	1.895	**0.175***
			Bark	21.32	0.0001
ER_F	a	F1	None (Hertz)	9.829	0.0001
			Bark	100.417	0.0001
		F2	None (Hertz)	0.624	**0.545***
			Bark	152.526	0.0001
JF_M	a	F1	None (Hertz)	1.603	0.203*
			Bark	188.648	0.0001
		F2	None (Hertz)	1.465	**0.254***
			Bark	16.416	0.0001
LS_M	a	F1	None (Hertz)	311.404	**0.091***
			Bark	104.406	0.0001
		F2	None (Hertz)	6.142	0.0001
			Bark	41.649	0.0001
OQ_M	a	F1	None (Hertz)	13.553	0.0001
			Bark	299.105	0.0001
		F2	None (Hertz)	16.76	0.0001
			Bark	119.552	0.0001

Speaker	Vowel	Formant	Normalizatin	Anova F(2,21)	p value
SB_M	a	F1	None (Hertz)	2.18	**0.138***
			Bark	12.429	0.0001
		F2	None (Hertz)	1.475	0.252
			Bark	8.211	0.002
SR_M	a	F1	None (Hertz)	0.122	**0.088***
			Bark	545.137	0.0001
		F2	None (Hertz)	0.542	**0.589***
			Bark	64.511	0.0001
VG_F	a	F1	None (Hertz)	6.041	0.0001
			Bark	313.315	0.0001
		F2	None (Hertz)	2.272	**0.128***
			Bark	10.898	0.001
VL_L	a	F1	None (Hertz)	16.09	0.0001
			Bark	409.9	0.001
		F2	None (Hertz)	3.177	**0.062***
			Bark	212.769	0.0001
AB_F	i	F1	None (Hertz)	12.82	0.0001
			Bark	340.616	0.0001
		F2	None (Hertz)	340.616	0.0001
			Bark	26.257	0.0001
AM_F	i	F1	None (Hertz)	31.617	0.0001
			Bark	207.533	0.0001
		F2	None (Hertz)	45.659	0.0001
			Bark	23.322	0.0001
CM_F	i	F1	None (Hertz)	0.105	**0.901***
			Bark	208.323	0.0001
		F2	None (Hertz)	2.7	**0.09***
			Bark	194.302	0.0001
DS_M	i	F1	None (Hertz)	1.057	**0.365***
			Bark	61.709	0.001
		F2	None (Hertz)	5.191	0.15
			Bark	10.351	0.001
ER_F	i	F1	None (Hertz)	13.096	0.0001
			Bark	362.458	0.001
		F2	None (Hertz)	5.117	0.015
			Bark	73.722	0.0001

Speaker	Vowel	Formant	Normalizatin	Anova F(2,21)	p value
JF_M	i	F1	None (Hertz)	17.193	0.0001
			Bark	215.032	0.0001
		F2	None (Hertz)	2.866	**0.079***
			Bark	13.935	0.0001
LS_M	i	F1	None (Hertz)	1.387	**0.272***
			Bark	80.012	0.0001
		F2	None (Hertz)	1.791	**0.191***
			Bark	7.915	0.003
OQ_M	i	F1	None (Hertz)	6.917	0.005
			Bark	100.186	0.0001
		F2	None (Hertz)	1.875	**0.178***
			Bark	10.374	0.001
SB_M	i	F1	None (Hertz)	16.002	0.001
			Bark	101.575	0.0001
		F2	None (Hertz)	8.377	0.002
			Bark	10.674	0.001
SR_M	i	F1	None (Hertz)	45.603	0.0001
			Bark	140.523	0.0001
		F2	None (Hertz)	13.499	0.001
			Bark	10.888	0.01
VG_F	i	F1	None (Hertz)	5.688	0.011
			Bark	82.799	0.0001
		F2	None (Hertz)	5.879	0.009
			Bark	34.256	0.0001
VL_L	i	F1	None (Hertz)	2.486	**0.107***
			Bark	74.345	0.0001
		F2	None (Hertz)	51.672	0.0001
			Bark	16.389	0.001
AB_F	u	F1	None (Hertz)	3.132	**0.064***
			Bark	552.616	0.0001
		F2	None (Hertz)	8.594	0.002
			Bark	163.334	0.0001
AM_F	u	F1	None (Hertz)	21.158	0.0001
			Bark	121.192	0.001
		F2	None (Hertz)	1.839	**0.184***
			Bark	8.544	0.002

Speaker	Vowel	Formant	Normalizatin	Anova F(2,21)	p value
CM_F	u	F1	None (Hertz)	0.122	**0.886***
			Bark	169.6	0.0001
		F2	None (Hertz)	0.13	**0.879***
			Bark	57.544	0.001
DS_M	u	F1	None (Hertz)	4.572	0.022
			Bark	214.468	0.0001
		F2	None (Hertz)	8.558	0.002
			Bark	37.126	0.0001
ER_F	u	F1	None (Hertz)	9.184	0.001
			Bark	162.399	0.0001
		F2	None (Hertz)	0.904	**0.42***
			Bark	10.452	0.001
JF_M	u	F1	None (Hertz)	8.228	0.002
			Bark	56.516	0.001
		F2	None (Hertz)	19.969	0.001
			Bark	8.173	0.02
LS_M	u	F1	None (Hertz)	2.434	**0.112***
			Bark	211.774	0.001
		F2	None (Hertz)	0.074	**0.929***
			Bark	4.327	0.027
OQ_M	u	F1	None (Hertz)	11.791	0.0001
			Bark	137.459	0.0001
		F2	None (Hertz)	48.9	**0.63***
			Bark	56.019	0.0001
SB_M	u	F1	None (Hertz)	14.302	0.0001
			Bark	421.666	0.0001
		F2	None (Hertz)	2.377	**0.117***
			Bark	11.651	0.0001
SR_M	u	F1	None (Hertz)	0.656	**0.529***
			Bark	86.624	0.0001
		F2	None (Hertz)	2.475	**0.108***
			Bark	28.585	0.0001
VG_F	u	F1	None (Hertz)	3.323	**0.056***
			Bark	196.262	0.0001
		F2	None (Hertz)	3.242	**0.059***
			Bark	10.899	0.001

Speaker	Vowel	Formant	Normalizatin	Anova F(2,21)	p value
VL_L	u	F1	None (Hertz)	3.191	**0.062***
			Bark	67.891	0.0001
		F2	None (Hertz)	2.565	**0.101***
			Bark	10.493	0.001