A PRELIMINARY INVESTIGATION OF VOCALIC REALIZATIONS OF HISPANIC MICRO-WORKERS FROM LATIN AMERICA

Ioana Vasilescu, Tubaro Yaru Wu, Lamel Lori, Torres Juana, Paola Cierpe

A PRELIMINARY INVESTIGATION OF VOCALIC REALIZATIONS OF HISPANIC MICRO-WORKERS FROM LATIN AMERICA

Ioana Vasilescu[1], Yaru Wu[2], Lori Lamel[1], Juana Torres Cierpe[3], Paola Tubaro [4]

[1]LISN, Univ. Paris-Saclay, Orsay, France,

[2]CRISCO/EA4255, Université de Caen Normandie, Caen, France,

[3]Télécom Paris, Institut Polytechnique de Paris, France,

[4]CNRS & CREST, UMR 9194, Palaiseau, France

{ioana.vasilescu, lori.lamel}@lisn.fr, yaru.wu@unicaen.fr, juana.torrescierpe@telecom-paris.fr, paola.tubaro@cnrs.fr

ABSTRACT

This paper focuses on socio-phonetic features of micro-workers supporting the data and annotation needs for artificial intelligence, from several Latin American countries. It represents a first attempt to document the linguistic identity of this emerging professional community. We raise the issue of a group of speakers that might not fall under classical sociolinguistic variables as they have the same activity without sharing a common workplace. Instead, they are spread throughout a large geographical area and likely to cover several regional varieties of Spanish. We make use of a corpus of interviews recorded as part of a larger sociological survey and focus on the acoustic properties of vowels as function of socio-demographic variables. We perform principal component analysis and hierarchical clustering to correlate formant values with socio-demographic variables. The preliminary results suggest vocalic realizations and reduction patterns correlate with level of education and social status.

Keywords: socio-phonetics, LA Spanish, vocalic quality, socio-demographic variables.

1. INTRODUCTION

The present work is a first attempt to document the linguistic characteristics of the emerging professional community of micro-workers supporting the data and annotation needs for artificial intelligence (AI). We exploit a corpus of recorded interviews as part of sociological surveys, to investigate patterns of spoken variation of AI micro-workers from several Latin America (henceforth LA) countries. "Microwork" encompasses a range of small, fragmented tasks performed remotely online by large numbers of providers [1]. It is a by-product of the "datafication" of our societies [2], whereby more data and increasing computational capacity have brought AI research to the fore and in turn, further raised the demand for ever-larger, ever-better data assets. AI micro-workers operate via online platforms and perform tasks as varied as annotating images, recording short sentences or correcting a speech recognition systems' output. Since Labov's founding work [3, 4], the socio-linguistic analysis described patterns of language variation of many languages and from many kinds of data. Therefore, today we know that different demographic and socio-economic groups deploy different linguistic behaviors and characteristics, as do the inhabitants of different geographic areas and particular populations over time. On this basis, it can be assumed that the larger the territory where a language is spoken, the greater the number of linguistic varieties.

This paper addresses the challenge of phonetic particularities of a group of speakers that may not fall under classical socio-linguistic variables. The speakers in our corpus are micro-workers recruited through an international platform open to Spanish speakers and they all declare Spanish as mother tongue. They perform similar tasks without sharing an identified workplace. According to the sociological survey our speakers are coming from different social backgrounds [5]. If some of them are in virtual contact *via* forums related to work on platforms, they are spread through a large geographical area as they are coming from 6 LA countries (Argentina, Colombia, Dominican Republic, Ecuador, Peru and Venezuela). They are thus likely to cover several regional varieties of LA Spanish. We focus here on the acoustic properties of vowels as function of socio-demographic variables.
The remainder of this paper is organized as follows. Section 2 provides a brief survey of patterns of segmental variation in Spanish and motivates the focus on vowels. Section 3 describes the corpus and methodology, followed by an analysis and preliminary results in Section 4. We conclude in Section 5.

2. SPANISH VOWELS VARIATION: A STATE OF THE ART REVIEW

Today Spanish is one of the top 5 most spoken languages in the world and mother tongue for nearly 500 million speakers[1]. With such a worldwide coverage, it benefits from an extensive body of work dedicated to patterns of variation at many linguistic levels and with respect to various socio-linguistic factors [6]. The reason we focus on segmental variation and, in particular, on the acoustic properties of vowels, is twofold. First, we want to contribute to the socio-linguistic study of variation in Spanish of an occupational group hitherto ignored by linguistic explorations but increasingly present in social science studies. State of the art of socio-linguistic perspective points out a range of factors responsible of vocalic variation in Spanish. Among them, the role of bilingualism and language contact has been often mentioned. Speech style is another factor, as well as gender, sexual orientation, social class, and social networks [7, 8]. Our data offer the opportunity to take into account other variables such as education level and social status based on the presence vs. absence of a main occupation, and the reported income.

Second, we aim at contributing to the growing work dedicated to vocalic variation in Spanish. Spanish vocalic system has been described as a particularly stable five-vowel system inherited from Latin. The vowels show more stability compared to consonants that seem to present much more variation especially between Peninsular and LA Spanish [9, 10, 11]. Despite this stability, increasing work is dedicated to patterns of dialectal and socio-linguistic variation of Spanish vowels which seem to be less stable than previously thought [12, 13, 14, 15]. Among the patterns of variation that affect Spanish vowels, reduction is often mentioned, even if other phenomena are also discussed such as vowel harmony [6]. Vocalic reduction leads to weakened vowels (eg. shortened, unstable and even deleted realizations) and is responsible for changes in vocalic quality (formants and voicing). Both lexical stress and word position play a role in vocalic reduction, at least as far as LA varieties are concerned [16, 6], whereas studies show that stress does not have a conventionalized effect on vowel quality in Iberic Spanish [17].

3. CORPUS AND METHODOLOGY

This section presents the data and methodology, and describes the main features of the corpus.

3.1. Corpus acquisition

The dialogues were acquired in the framework of the project TRIA ("The Real voices of Artificial Intelligence"), an interdisciplinary project involving sociologists, economists and linguists from Paris-Saclay University and Institut Polytechnique de Paris. The aim of the project was to document the lived experiences and identity of platform workers from different countries. In particular it focused on the digital labour in Spanish speaking countries, increasingly known as one of the most important reservoirs of micro-work in Europe and beyond[2]. TRIA is a two-part corpus:

1. Written: online questionnaires to document the socio-economic profiles of micro-workers. The surveys consist of over 100 questions covering a wide range of socio-demographic information from gender, age and level of education to patterns of their online earning activities (eg. tasks, platforms used etc.) with the aim of defining their socio-economic profile as comprehensively as possible.
2. Spoken: in-depth interviews with a sub-sample of the same micro-workers, with the objective of corroborating questionnaires. The interviews are scripted to cover the following topics: career path, arrival on the micro-work market, the micro-working tasks and their place in the life of the micro-worker (i.e. the balance between micro-working and everyday life). The wording of the questions as well as the time allocated to each question varies from one interview to another.

The available data correspond to interviews with 57 micro-workers, resulting in approximately 40 hours of speech. For this preliminary work we have processed and we report on 14 hours of speech, corresponding to 18 interviews conducted with 13 male and 5 female micro-workers. They cover the 6 countries mentioned in Section 1: Argentina (4), Colombia (3), Ecuador (2), Peru (1), Dominican Rep. (1), Venezuela (6). Among them 3 micro-workers have a migration history (Spain-Ecuador, Venezuela-Argentina and Venezuela-Colombia). The interviewers are 3 female sociologists (2 Spanish natives, 1 bilingual Italian-Spanish). All the interviews were conducted over videocalls due to the pandemic during the year 2020-2021. Volunteers received compensation for their participation.

3.2. Methodology

Interviews were manually transcribed by a native speaker of Spanish (the fourth author). Then

Corpus	Counts
#duration	14h
#duration speech	11h (80%)
Mean dur./interview	48min.
Mean dur./speaker turn	23min.
#words	90632
#tokens	6441
#tokens micro-w.	5553

Table 1: General description of the TRIA corpus.

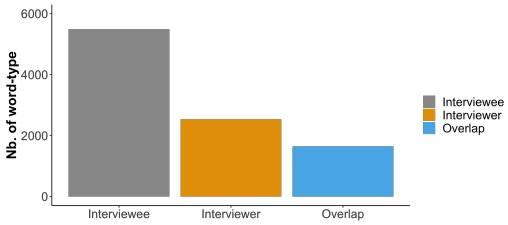

Figure 1: Word-types used by interviewee (gray), interviewer (yellow) and overlapped speech (blue).

the language-specific speech recognition system described in [18] was used to carry out a forced alignment of the speech with the manual transcription. The acoustic models were trained on data from both Peninsular and LA Spanish.

The novelty of our approach is that we take advantage of the socio-demographic variables available through the written survey to enrich the speech with meta-data corresponding to the profile of the micro-workers. Thereafter a mapping is realized between each speaker turn and the corresponding variables. The selected labels, which will later serve as factors are the following: Gender (male, female), Country of residence (Argentina, Colombia, Dominican Republic, Ecuador, Peru, Venezuela), Education level (Secondary (high-school), Post-secondary, Short Higher Education (up to 2 years), Bachelor's degree (3 years), Master's or higher (5 years), working situation (unemployed, family assistance, fixed-term contract, open-term contract, employer, independent, student), Primary income (present, absent).

We focus here on the acoustic quality of the vowels and used PRAAT [19] to extract formants values for the first four formants (F1, F2, F3 and F4). First, we used the formants values to perform a principal component analysis (PCA) in R [20]. Thereafter we applied hierarchical clustering on principal components (HCPC) on the result of the PCA. In order to illustrate individual entries of HCPC result, we calculated the mean value of

Vowel	a	e	i	o	u
occ.	26450	31851	12551	22917	6163

Table 2: Nbr. of lexical vowels used in the study.

F1, F2, F3, F4 for each factor (each level of the meta-data as mentioned above). Second, we used the measurements of the first and second formants (F1, F2) to generate the vocalic space according to selected factors (see Section 4). Three values are extracted for each vowel: one at the beginning, one in the middle and one at the end of the segment.

3.3. Corpus features

The general characteristics of the data used in the study are summarized in Table 1. Figure 1 shows that interviewees' (micro-workers) speaking time is significantly longer compared to the interviewers (sociologists), suggesting a rather spontaneous character of the exchanges. The presence of overlapping speech as well as a ratio of verbal vs non-verbal speech events (here empty and filled pauses) similar to other Spanish spontaneous corpora previously analyzed [18] suggest that TRIA has the features of a spontaneous corpus (Figure 1). From there we hypothesise that phenomena specific to spontaneous speech may occur such as reductions and/or non-canonical realizations of segments. With respect to the vocalic system, previous investigations dedicated to acoustic quality of the Spanish vowels as function of the speech style underlined universal (i.e. language-independent) trends such as the tendency to centralization of the vocalic system [21]. In addition, studies also pointed out the propensity to schwa-like realizations of Spanish vowels modulated by lexical stress, the unstressed position triggering more centralized realizations as they correspond to shorter realizations compared to stressed vowels [22].

4. PRELIMINARY RESULTS

We focus on the acoustic quality of monophthong vowels regardless of the lexical stress: near 100k occurences are used for the analysis (Table 2). The acoustic quality of the Spanish vowels is examined with respect to socio-demographic meta-data used as factors in the two following procedures. First, mean values of F1, F2, F3 and F4 of the focal vowels /i-a-u/ are used to perform Principal Component Analysis (PCA), followed by HCPC (hierarchical clustering on principal components), in order to have an estimate of the vowels co-variation according to socio-demographic labels (i.e. Gender, Country of Residence, Diploma, Professional situation, +/- Income). Figures 2, 3 and 4 show the co-

variation for each vowel. The two axes represented correspond to the first two dimensions of the PCA models. Note that the dimensions are not directly formants, but the two dimensions that model the best the dataset. Dispersion show that Gender is predictably responsible for clustering and for all vowels (Y-axis). However, they suggest that other variables play a role in the clusters' composition, such as level of education and socio-economic situation in a broad sense (e.g. the main occupation, the reported income). In particular, X-axis separates values corresponding to lower vs. higher level of education in particular for the vowel /a/.

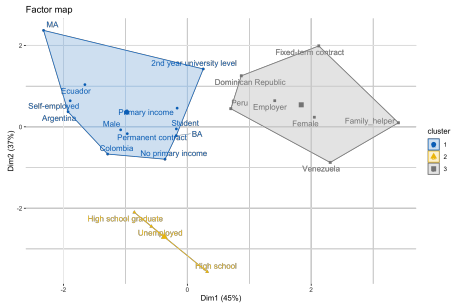

Figure 2: Visualization of categories of each factor based on the PCA map and HCPC : [a].

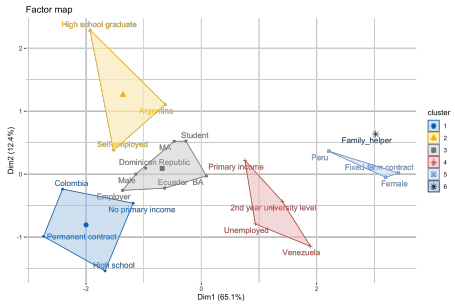

Figure 3: Visualization of categories of each factor based on the PCA map and HCPC : [i].

Figure 4: Visualization of categories of each factor based on the PCA map and HCPC : [u].

Second, we used the more represented variables in the dataset, that is Gender and reported Income, to represent the 5 Spanish vowels in the F1/F2 space. Table 3 groups the speakers according to the source of income and, as a result, the main or complementary character of the micro-work (+/-Income), and underlines the correlation with Gender. Figure 5 suggests the tendency to centralization potentially correlated to selected variables, e.g. male and female speakers for which micro-work is the main source of income produce the more centralized and thus potentially reduced vowels. Such profiles also correlate with lower level of education suggesting diaphasic variation as function of the social status.

Income	Male	Female
+Income	7	5
-Income	7	0

Table 3: Data distribution as function of Gender and reported Income.

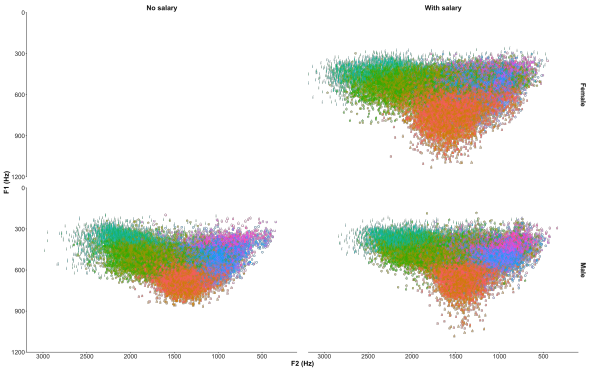

Figure 5: Vocalic dispersion according to Gender and reported Income.

5. CONCLUSION

We proposed a preliminary analysis of the acoustic quality of Spanish vowels from a corpus of interviews made by sociologists with micro-workers from Latin America. Measurements of the first four formants of the focal vowels were used to conduct principal component analysis and hierarchical clustering with respect to the socio-demographic labels. The most represented labels (Gender, Income) were used as factors to compare several vocalic dispersions as a function of F1/F2. The results suggest that educational profile and the socio-economic situation influence in a measurable manner the vocalic variation. This multidisciplinary research is in progress: we will both enrich the corpus with more data and add other parameters of variation to draw an in-depth picture of the micro-workers from Latin America and beyond.

6. ACKNOWLEDGEMENTS

This research was supported by the MSH Paris-Saclay grant "TRIA (Le TRavail de l'Intelligence Artificielle)" awarded to Paola Tubaro.

7. REFERENCES

[1] P. Tubaro and A. A. Casilli, "Micro-work, artificial intelligence and the automotive industry," *Journal of Industrial and Business Economics = Economia e politica industriale*, vol. 46, p. 333–345, Jun. 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02148979

[2] G. Bastin and P. Tubaro, "Le moment big data des sciences sociales," *Revue française de sociologie*, vol. 59, no. 3, pp. 375–394, Sep. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01885416

[3] W. Labov, "The study of language in its social context," in *Volume 1 Basic concepts, theories and problems: alternative approaches*, J. A. Fishman, Ed. Berlin, Boston: De Gruyter Mouton, 1971, pp. 152–216.

[4] W. Labov, I. Rosenfelder, and J. Fruehwald, "One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis," *Language*, vol. 89, pp. 30–65, 2013.

[5] P. Tubaro, "Learners in the loop: hidden human skills in machine intelligence," *Sociologia del Lavoro*, vol. 163, pp. 110–129, 2022. [Online]. Available: https://hal.inria.fr/hal-03787017

[6] J. M. Lipski, *Geographical and Social Varieties of Spanish: An Overview*. John Wiley Sons, Ltd, 2012, ch. 1, pp. 1–26. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118228098.ch1

[7] M. de la Fuente Iglesias and S. Pérez Castillejo, "Spanish mid vowels as sociolinguistic variables in galicia," *Spanish in Context*, vol. 17, no. 3, pp. 464–487, 2020. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/sic.18027.igl

[8] R. Ronquest, *Vowels*, ser. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2018, p. 145–164.

[9] J. Hualde, *The Sounds of Spanish*. Cambridge University Press, 2005. [Online]. Available: https://books.google.be/books?id=vI-wswEACAAJ

[10] P. Delattre, "An acoustic and articulatory study of vowel reduction in four languages." *Iral-international Review of Applied Linguistics in Language Teaching*, vol. 7, pp. 295–326, 1969.

[11] T. Navarro Tomás, *Manual de pronunciación española /*, 13th ed. Madrid:: Monteverde,, 1967.

[12] J. I. Hualde, M. Simonet, and F. Torreira, "Postlexical contraction of nonhigh vowels in spanish," *Lingua*, vol. 118, no. 12, pp. 1906–1925, 2008, studies on the Phonetics and Phonology of Glides. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0024384107001787

[13] M. F. Dabkowski, "Acoustic properties of mexico city spanish vowel weakening," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3578–3578, 2017. [Online]. Available: https://doi.org/10.1121/1.4987616

[14] S. M. Alvord and B. Rogers, "Miami-cuban spanish vowels in contact," *Sociolinguistic Studies*, vol. 8, no. 1, p. 139–170, Jul. 2014. [Online]. Available: https://journal.equinoxpub.com/SS/article/view/7300

[15] R. Ronquest, "Stylistic variation in heritage spanish vowel production," *Heritage Language Journal*, vol. 13, no. 2, pp. 275 – 298, 2016. [Online]. Available: https://brill.com/view/journals/hlj/13/2/article-p275_9.xml

[16] J. Hualde and I. Chitoran, "Surface sound and underlying structure: The phonetics-phonology interface in romance languages," in *Manual of grammatical interfaces in Romance*, S. Fisher and C. Gabriel, Eds. De Moution Gruyter, 2016, vol. 10, pp. 23–44.

[17] M. Nadeu, "Stress- and speech rate-induced vowel quality variation in catalan and spanish," *Journal of Phonetics*, vol. 46, pp. 1–22, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447014000540

[18] I. Vasilescu, N. Hernandez, B. Vieru, and L. Lamel, "Exploring temporal reduction in dialectal spanish: A large-scale study of lenition of voiced stops and coda-s," in *Proc. Interspeech 2018*, 2018, pp. 2728–2732. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1256

[19] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[21] B. Harmegnies and D. Poch-Olivé, "A study of style-induced vowel variability: Laboratory versus spontaneous speech in spanish," *Speech Communication*, vol. 11, no. 4, pp. 429–437, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016763939290048C

[22] F. Santiago and P. Mairano, "The role of lexical stress on vowel space and duration in two varieties of Spanish," Proceedings of the 9th International Conference on Speech Prosody, pp. 453–457, 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01717914

[1] https://www.exteriores.gob.es/en/PoliticaExterior/Paginas/ElEspanolEnElMundo.aspx

[2] The study protocol was approved by the Data Protection Officer of CNRS under GDPR