Using Speech Technology to test Theories of Phonetic and Phonological Typology

Anisia Popescu, Lori Lamel, Ioana Vasilescu
LISN, CNRS, Université Paris Saclay
anisia.popescu@universite-paris-saclay.fr

Abstract

The present paper uses speech technology derived tools and methodologies to test theories about phonetic typology. We specifically look at how the two-way laryngeal contrast (voiced /b, d, g, v, z/ vs. voiceless /p, t, k, f, s/ obstruents) is implemented in European Portuguese, a language that has been suggested to exhibit a different voicing system than its sister Romance languages, more similar to the one found for Germanic languages. A large European Portuguese corpus was force aligned using (1) different combinations of parallel Portuguese (original), Italian (Romance language) and German (Germanic language) acoustic phone models and letting an ASR system choose the best fitting one, and (2) pronunciation variants (/b, d, g, v, z/ produced as either [b, d, g, v, z] or [p, t, k, f, s]) for obstruent consonants. Results support previous accounts in the literature that European Portuguese is diverging from the traditional voicing system known for Romance language, towards a hybrid system where stops and fricatives are specified for different voicing features.

Keywords: corpus-based linguistics, acoustic models, forced alignment, pronunciation variants, voicing systems, laryngeal contrast

1. Introduction

In recent years methodologies initially developed for automatic speech recognition (ASR) and natural language processing (NLP) have been repurposed and applied in other domains as varied as healthcare, administration, social sciences or education. Linguistic research is another domain that made use of speech technology methodologies: Coleman et al. (2016) looked at nasal place assimilation; Renwick et al. (2016) quantified phonological contrast between vowels; Yuan and Liberman (2011) investigated lateral allophones. The present study focuses on phonetic and phonological typology. In the last decade several laboratory phonology studies (Pape and Jesus, 2011, 2015; Ramsammy and Strycharczuk, 2016; Jesus and Costa, 2020) have suggested European Portuguese (henceforth **EP**) stands apart from its Romance family relatives when it comes to obstruent voicing. Small scale acoustic studies show that obstruent voicing profiles of EP obstruents resemble those of German, a Germanic language, rather than those of Italian, a related Romance language. Romance languages are generally known to be "true voicing" languages, which mark the laryngeal contrast (voiced vs. voiceless obstruents) through the [voice] feature (i.e., there is actual vocal cord vibration during the production of /b,d,g, v, z, ʒ/ obstruent series). Germanic languages are usually considered to be "aspirating" languages, marking the contrast through the feature [spread glottis] (i.e., there is no actual vocal cord vibration for /b,d,g, v, z, ʒ/ except in intersonorant position). Ramsammy and Strycharczuk (2016) suggest EP, a Romance language, exhibits a hybrid voicing system where [spread glottis] is the contrast feature for fricatives whereas [voice] better handles the contrast in the case of stops. Looking at aerodynamic measurements Jesus and Costa (2020) suggest voicing in stops, not only fricatives, is also better described using the [spread glottis] feature. Both studies use small scale acoustic/aerodynamic measurements to infer their results. The present paper further investigates the voicing system in EP using two methods derived from speech technology: (1) allowing an ASR system to choose between a combination of different language acoustic phone models when force aligning EP speech data and (2) force aligning EP speech data with voice/voiceless pronunciation variants for obstruent consonants. In order to test whether this methodology replicates results found in laboratory studies we compare EP with Italian and German, the original languages used in (Pape and Jesus, 2015, 2011). We aim to answer the following research questions:

- Can speech technology methods be used to test theories of phonetic and phonological typology?
- Is EP a true voicing language like its genetically related Romance languages, or did it shift towards an "aspirating" language voicing system?
- Is voicing in EP obstruents more similar to Italian, a sister Romance language, or to German, a Germanic language?

LREC-COLING 2024, pages 17321–17325
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

2. Methodology

To answer these questions we looked at an EP corpus consisting of 114 hours of mostly standard dialectal broadcast news speech from radio and TV shows. We made use of existing data by exploiting corpora previously acquired from the LDC, ELRA and several international projects. The corpus was forced aligned matching phones to their orthographic transcription using a trained acoustic model and pronunciation dictionary for Portuguese. The output of the forced alignment is a sequence of phones with labels predicted by the Portuguese language dictionary.

The first method to answer our research questions, involved adding two different sets of obstruent (stop and fricative) phone models, one for Italian and one for German, in parallel to the original EP obstruent phone model. The acoustic phone models for all other phonemes were kept in their original EP form. The acoustic phone models for all three considered languages (EP, Italian and German) were trained on similar type of data (broadcast news speech i.e., (semi)prompted speech), similar amounts of time (roughly 100 hours), and comparable number of word tokens and types (EP: 1.1. million words tokens and 46k word types; Italian: 1.8 million word tokens, 58.8k word types; German: 1.8 million word tokens and 90k word types). Acoustic models for each language are speaker-, context- and word-position-independent monophone models. Each phone model is a 3-state left-to-right continuous density HMM with Gaussian mixtures with up to 32 Gaussians per states. Silences are modeled by a single state with 256 Gaussians. The same acoustic parametrization was used for all phone models: cepstral - PLP (Hermansky, 1990) and pitch (F0) features. A similar procedure to the one used is described in Lamel et al. (2011). Presenting the recognition system with combinations of different language acoustic models allows us to force the system to choose the best fitting model (either the original EP, the Italian or the German) for each individual phonemically voiced stop and fricative in the corpus (illustrated in 1 for the Portuguese word *dado* [dadu] 'given').

Table 1 shows the counts of phonemically voiced stops and fricatives in the corpus. The postalveolar /ʒ/ was left out of the analysis since it is not included in the Italian phoneme inventory and it appears only in loanwords in German. Two sets of language acoustic model combinations were used for the study: (1) a three way choice of acoustic models between EP, German and Italian, and (2) a binary choice between Italian and German acoustic obstruent models (EP obstruent models were no longer available to the system in this case). Each of the two combinations will be describes in different sections. We predict that if obstruent voicing in EP is indeed more similar to German, as shown by (Pape and Jesus, 2015), we would expect the system to choose the German acoustic models to a greater extent than it does the Italian obstruent models. If however, the opposite is true (i.e, the voicing system in EP is consistent to the one found in Romance languages) we would expect the system to prefer the Italian obstruent acoustic models.

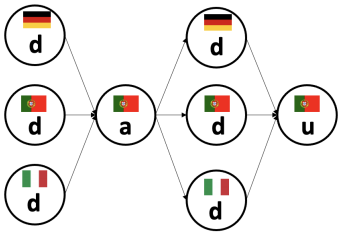

Figure 1: Combination of three acoustic models (EP, Italian and German - represented by flags) for the stop /d/ in the Portuguese word *dado* [dadu] 'given'.

Manner	Consonant	Count
stop	/b/	43,400
stop	/d/	225,705
stop	/g/	42,124
fricative	/v/	67,742
fricative	/z/	47,306

Table 1: Counts of phonemically voiced stops /b, d, g/ and fricatives /v, z/ in the corpus.

A second method was used to answer our research questions: forced alignment with pronunciation variants. The method was first introduced by Hallé and Adda-Decker (2011) to study voicing assimilation in French and has been successfully used to investigate non-canonical voicing in Romance languages is several studies (Popescu et al., 2023; Wu et al., 2022; Hutin et al., 2022; Vasilescu et al., 2020). It implies using one language acoustic model at a time (not in parallel) with the addition of pronunciation variants. The Portuguese language dictionary was enriched with pronunciation variants for obstruent voicing. For example the Portuguese word *dado* [dadu] - 'given' had four possible pronunciations [dadu], [datu], [tadu] or [tatu]. The same applies for fricatives (/vir/ 'come' could be detected as either the original voiced [viʁ] or the devoiced [fiʁ]). The system then has to chose which phone model (phonetically voiced or voiceless) best fits the data. Results of this method will be detailed in section 5.

3. Three-way choice of acoustic models - Portuguese, Italian and German

In this experiment, for each phonemically voiced obstruent /b,d,g, v, z/ the system was presented with three different language phone models (original EP, Italian and German). The system then had to choose which of the three phone models best fits the acoustic signal corresponding to the EP obstruents. Figure 2 shows the counts (y-axis) and percentages of selected phone models per language as a function of place of articulation. Note that the scales are different between stops and fricatives: stops are more frequent (coronal /d/ in particular).

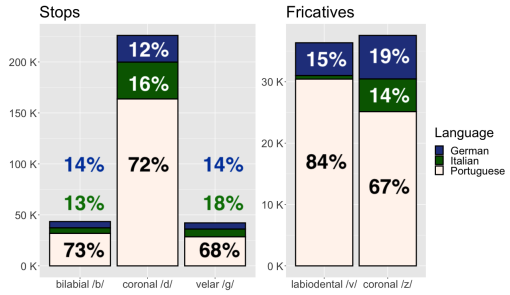

Figure 2: Counts and percentages of phone occurrences aligned with one of three acoustic phone models (Portuguese in white, German in blue or Italian in green) for stops (left) and fricatives (right).

As expected the original Portuguese model is preferred (in 72% of cases for stops and 75% of cases for fricatives). In the rest of cases the system preferred either the German or the Italian phone models. For stops percentages for Italian and German phone models are similar, with German phone models being marginally preferred over the Italian ones (14% vs. 13%). For coronals and bilabials Italian phone models are preferred. For fricatives for both places of articulation the German phone model is preferred.

4. Two-way choice of acoustic models - Italian and German

In this second experiment, only two obstruent acoustic phone models were presented to the system to chose from: the Italian and the German ones. The original Portuguese phone model was no longer an option. Figure 3 shows the counts and percentages of obstruent occurrences aligned with either an Italian or German obstruent acoustic model.

Results show the same patterns as in the previous section. For stops the Italian phone models are

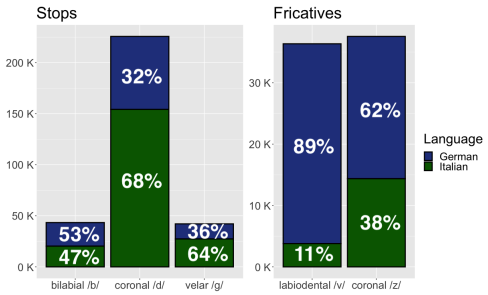

Figure 3: Counts and percentages of phone occurences aligned with one of two acoustic phone models (Italian in green or German in blue) for stops (left) and fricatives (right).

preferred (68% of cases for coronals and 64% of cases for velars) except for bilabials for which the German phone models are preferred (German 53% vs. Italian 47%). For fricatives for both places of articulation (labiodental and coronal) the German phone models are markedly preferred.

In summary, the results presented in Sections 3 and 4 confirm our predictions only in the case of fricatives (i.e, EP fricatives are similar to those of German) but not in the case of stops (with only bilabial stops presenting the predicted patterns). So far results point towards there being a hybrid voicing system for obstruent stops in EP, as suggested by Ramsammy and Strycharczuk (2016). In the next section we present results of the forced alignment with pronunciation variants.

5. Forced alignment with pronunciation variants

In this third experiment, we enriched the Portuguese pronunciation dictionary with pronunciation variants for obstruent voicing (voiced obstruents /b, d, g, v, z/ could be detected by the system as either voiced [b, d, g, v, z] or voiceless [p, t, k, f, s]) allowing the system to choose the best fitting phone model (voiced of voiceless) for each phonemically voiced obstruent in our corpus. Figure 4 shows the waveform, spectrogram and alignment of the word *dado* [dadu] 'given'. The first /d/ consonant is detected by the system as a voiceless [t], the second as voiced [d].

We ran the alignment twice, once with Italian and once with German voiced/voiceless obstruent phone models. For example, when using the German obstruent acoustic models, if a Portuguese phonemically voiced stop /d/ better matched the German phone model [d], the output of the system for that stop would be [d]. If, however, the voiceless acoustic model better fit the data, the output would be a voiceless [t]. This second method differs from the one presented in the previous two sections in

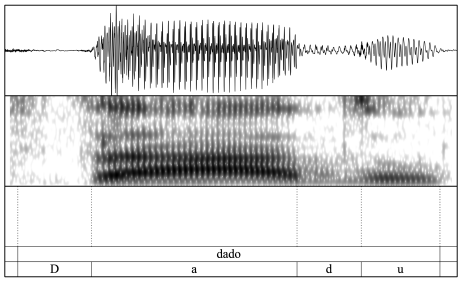

Figure 4: Waveform and spectrogram of the Portuguese word *dado* [dadu] 'given'. The first /d/ (labeled D) is detected as a voiceless [t] and the second /d/ (labeled d) is detected as a voiced [d] by the system.

that it allows us to test the similarity/differences between EP and Italian/German from a different perspective. Based on previous acoustic studies, we know that the voicing profiles (probability of voicing in 10 different points throughout the consonant) differ based on language: while Italian voicing probability remains close to one throughout the obstruent, the voicing probablity drops after 30% of the obstruent for both EP and German (Pape and Jesus, 2015). This suggest that both German and EP exhibit partial devoicing of the obstruent and Italian does not. We therefore expect the system to chose more voiceless variants (higher percentages of voiceless variants) when force aligning the data with Italian obstruent phone models (i.e., the partial devoicing in Portuguese obstruents would be interpreted as voicelessness by Italian acoustic models). Figure 5 shows the percentages of phonetically voiceless variants (color shades: blue for German, green for Italian) detected by the system when using the Italian or German voiced-voiceless acoustic model variants. Grey shades correspond to the phonetically voiced variants identified by the system.

Results confirm our predictions for fricatives - the Italian voiceless variants are selected at a greater extend than the German ones (Italian 90% vs. German 72% for the labiodental /v/, and Italian 57% vs. German 30% for the coronal /z/). For stops the results are more ambiguous: Italian voiceless variants are selected with a higher degree for bilabias and velars but not for coronals. These results go against our predictions and against the patterns found in Sections 3 and 4, where coronals patterned with velars, differently than bilabials (here the coronals pattern differently).

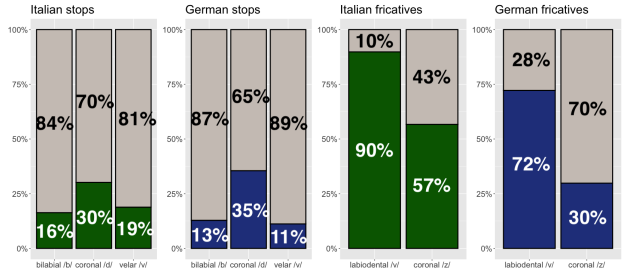

Figure 5: Percentages of phone occurrences aligned as with either the voiceless (German in blue, Italian in green) or voiced (grey shade for both languages) per language for stops (left) and fricatives (right)

6. Limitations

We proposed two methodologies that are not a direct replication of the original acoustic studies: while the acoustic studies relied on targeted acoustic (Praat's (Boersma and Weenink, 2019) autocorrelation (AC) pitch extraction, VOT, intensity or center of gravity) and/or aerodynamic (oral airflow and electroglottographic) measures, the present study relies exclusively on the trained acoustic models, which incorporate a set of acoustic features including log energy, pitch and 12 cepstrum coefficients. The difference in methodologies could be the reason for the diverging results found for stops. For a better understanding of the factors at play, an acoustic analysis of a subset of the present data is needed. Future studies should also include other acoustic correlates of voicing, such as phonetic/phonological context and prosodic information. Prosody is especially important since stress patterns in EP have also been shown to differ from other Romance languages which are syllabe-timed: the language is believed to be stress-timed (Cruz-Ferreira, 1999) like Germanic languages, or partially stress- and syllable-timed (Frota and Vigário, 2006).

7. Conclusion

The present paper tested theories of phonetic typology derived from small scale laboratory studies using two methodologies stemming from speech technology: (1) using different language combinations of acoustic models when force aligning data and (2) forced alignment with pronunciation variants. In particular we asked whether the EP obstruent voicing system is different than the one of its genetically related Romance languages, and more similar to the one found for Germanic languages. We find that there is a clear pattern of EP fricative voicing shifting away from the Romance language voicing system: EP fricatives are more similar to

those of German, than to those of Italian. This suggests [spread glottis] is the active feature for EP fricative voicing. For stops, the results are less conclusive. Place of articulation seems to play a role for stops, and we find different results depending on the used method. However, for both methods, two out of three stop consonants are more similar to Italian than to German (i.e., opposite pattern than the one found for fricatives). We take this as an indication that EP exhibits different voicing systems for stops and fricatives supporting Ramsammy and Strycharczuk (2016)'s account of EP having a hybrid voicing system. The results also support the repurposing of speech technology methodologies for studies in other fields such as theoretical and experimental linguistics.

P. Boersma and D. Weenink. 2019. Praat: doing phonetics by computer. *Computer program*.

J. Coleman, M.E.L Renwick, and R.A.M. Temple. 2016. Probabilistic underspecification in nasal place assimilation. *Phonology*, 33(3):425–458.

M Cruz-Ferreira. 1999. *Portuguese (European). Handbook of the International Phonetic Association: A guide to the use of the international phonetic alphabet*. Cambridge University Press, Cambridge.

S Frota and M Vigário. 2006. On the correlates of rhythmic distinctions: The European/Brazilian Portuguese case. *Probus*, 13(2):247–275.

P. Hallé and M. Adda-Decker. 2011. Bayesian framework for voicing alternation and assimilation studies on large corpora in french. In *In Proceedings of the 15th International Congress of Phonetic Sciences*, Saarbrücken, Austria.

H. Hermansky. 1990. Perceptual linear prediction (plp) analysis for speech. *J. Acoust. Soc. Amer.*, 87.

M. Hutin, M. Adda-Decker, L. Lamel, and I. Vasilescu. 2022. When phonetics meets morphology: Intervocalic voicing within and across words in Romance languages. In *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, pages 3438–3442, Incheon, Korea.

L.M. Jesus and M. Costa. 2020. The aerodynamics of voiced stop closures. *EURASIP Journal on Audio, Speech and Music Processing*, 2.

L. Lamel, S. Courcinous, J. Despres, J.L. Gauvain, Y. Josse, K. Kilgour, F. Kraft, V.B. Le, H. Ney, M. Nußbaum-Thom, I. Oparin, T. Schlippe, R. Schlüter, T. Schultz, T. Fraga da Silva, S. Stüker, M. Sundermeyer, B. Vieru, N.T. Vu, A. Waibel, and C. Woehrling. 2011. Speech recognition for machine translation in quaero. In *Proceedings of the 8th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 121–128, San Francisco, California.

D. Pape and L. M. Jesus. 2011. Devoicing of phonologically voiced obstruents: is european portuguese different from other romance languages. In *Proceedings of the 17th International Congres of Phonetic Sciences (ICPhS2011)*, pages 1566–1569, Hong Kong, China.

D. Pape and L.M. Jesus. 2015. Stop and fricative devoicing in euorpean portuguese, italian and german. *Language and Speech*, 58(2):224–245.

A. Popescu, M. Hutin, I. Vasilescu, L. Lamel, and M. Adda-Decker. 2023. Stop devoicing and place of articulation: A cross-linguistic study using large-scale corpora. In *In Proceedings of the 20th International Congress of Phonetic Sciences*, pages 3186–3190, Prague, Czech Reublic.

M. Ramsammy and P. Strycharczuk. 2016. From phonetic enhancement to phonological underspecification: hybrid voicing contrast in european portuguese. *Papers in Historical Phonology*, 1.

E. L. Renwick, I. Vasilescu ad C. Dutrey, L. Lamel, and B. Vieru. 2016. Martinal contrast among romanan vowels: Evidence from asr and functional load. In *Proceedings of the Interspeech 2016*, pages 2433–2437, San Frnacisco, US.

I. Vasilescu, Y. Wu, A. Jatteau, M. Adda-Decker, and L. Lamel. 2020. Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes. *TAL*, (61):11–36.

Y. Wu, M. Hutin, I. Vasilescu, L. Lamel, and M. Adda-Decker. 2022. Extracting linguistic knowledge from speech: A study of stop realization in 5 Romance languages. In *LREC*, pages 3257–3263, Marseille, France.

J. Yuan and . Liberman. 2011. /l/ variation in american english: A corpus approach. *Journal of Speech Sciences*, (2):35–46.