DERIVING MANNER OF ARTICULATION CLASSES FROM PHONEME CO-OCCURRENCE FREQUENCIES

Eleonora C. Albano

Laboratório de Fonética e Psicolinguística, University of Campinas, São Paulo, Brazil
albano@unicamp.br

ABSTRACT

This paper looks into the possibility of deriving manner of articulation classes from phoneme co-occurrence frequencies.

We show that distance measures based on phoneme co-occurrence frequencies group abstract manner of articulation classes that participate in phonological processes in spite of high acoustic and articulatory variability, e. g: obstruents, liquids, approximants, etc.

The test language is Brazilian Portuguese, as represented by two large phonemicized corpora. Phonetically meaningful groups of phonemes emerge from the statistical analysis of co-occurrence frequencies of types and tokens of phoneme pairs. The techniques employed are cluster analysis and multidimensional scaling.

Results show consistent groupings across corpora, counts, and statistical techniques. Besides consonants and vowels, co-occurrence frequencies satisfactorily separate the major classes, as well as some smaller manner of articulation subclasses.

Keywords: classification, manner of articulation, co-occurrence frequencies, multivariate statistics.

1. INTRODUCTION

Lack of acoustic invariance challenges direct induction of ordinary, lower-level phonetic classes such as dentals, velars, laterals, rhotics, etc. from the speech signal [1]. These classes also lack articulatory invariance [2]. Invariance is still more unlikely at higher-levels, such as the so-called major classes, e. g: obstruents, nasals, liquids, glides [3].

However, such abstract labels may act as:

- Targets of phonological processes, e.g., *liquid* devoicing is attested in English and Turkish among other languages [4];
- Triggers of phonological processes, e.g., pre-*sonorant* voicing is attested in Spanish, Dutch, and Slovak, among other languages [5].

It seems, therefore, that such higher-level information is in fact relevant to speakers/hearers. It is thus reasonable to ask whether there are other clues in the speech signal that might support the induction of abstract phonetic classes.

Phonotactics has been helpful in addressing another serious challenge to induction, namely, word segmentation. There is evidence that:

- Infants are sensitive to phonotactic probabilities [6];
- Phonotactically-based computer simulations succeed in extracting words from unsegmented running text [7].

Nevertheless, to the best of our knowledge, phonotactics remains practically unexplored as a bootstrapping mechanism for abstract phonetic classes. This paper is an attempt to clear this ground.

2. AIMS

- To show that co-occurrence frequencies of phonemes belonging to the same manner of articulation class tend to be highly correlated.
- To derive statistical groupings of phonemes from frequency-based distance measures, without recourse to any coding other than phonemization.
- To use multivariate exploratory statistical techniques to pave the way for the exploration of more powerful statistical classifiers to model induction of phonetic structure.

3. HYPOTHESES

- Phonetic structure grounded solely on probabilistic phonotactics will emerge from counts of both types and tokens of phoneme pairs in different corpora of the same language.
- Most groupings will involve manner of articulation classes, especially those directly related to sonority.

4. METHODOLOGY

Brazilian Portuguese (henceforth BP) is an appropriate language for a preliminary test of the above hypotheses because its syllable structure is relatively simple and faithful to the sonority principle. As in other Romance languages, onset clusters only allow for liquids in C_2, and coda clusters constitute a very limited set involving sonorants plus /s/. If sonority does indeed play a role in constraining phoneme combinations within and across syllables,

this should be clearly visible in BP. Evidently, such a prediction can only extend to languages with a similar syllable structure, but, if upheld, it will lead, anyway, to questions about languages with more complex phonotactics.

In addition, BP offers: (1) availability of large public corpora; (2) availability of automatic phonemization tools; (3) relative consensus on phonemic analysis, except for the question of biphonemic or monophonemic nasal vowels [8] – on which, however, it is easy to take a clear stand.

4.1. Corpora

Given the abstract nature of phonemic transcription, oral and written language should behave similarly, as their vocabularies, albeit different, follow the same syllable structure principles.

We have selected, therefore, the following two large corpora, available in the internet:

- LAEL-FALA (henceforth LAEL): a 44,967 word-type oral corpus, with 2,855,106 tokens, put together by the graduate program in applied linguistics of the Pontifical Catholic University of São Paulo, consisting of orthographic transcriptions of lectures, interviews and informal conversations [9].
- CETEN-FOLHA (henceforth CETEN): a 203,294 word-type corpus, with 22,600,865 tokens, put together by the Interinstitutional Centre for Computational Linguistics of the University of São Paulo at São Carlos, consisting of a full year edition (1994) of the newspaper Folha de São Paulo [10].

4.2. Treatment

4.2.1. Orthography-to-phoneme conversion

As both corpora were available only in conventional orthography, it was necessary to convert graphemes into phonemes. To this end, we used software created at our own laboratory.

4.2.2 Phoneme inventory

The phonemic analysis performed is minimalist in that it aims at minimal consonant and vowel inventories. To accomplish this, we took a rather conservative, structuralist approach:

- Allophony was resolved by having the least marked member represent the entire class (e.g., [s, z] > /s/).
- Glides, whether nasal or oral, were treated as vowels.
- Nasal vowels and glides were treated as biphonemic, e.g., [õ] > /on/, [ɐ̃õ] > /aʊn/.

This analysis yields an inventory of 19 consonants and 7 vowels, as follows:

- Consonants: p, b, f, v, m, t, d, s, z, n, l, r, ʃ, ʒ, ɲ, ʎ, k, g, R.
- Vowels: i, e, ɛ, a, ɔ, o, u.

In the same way as European Portuguese, BP has an intervocalic contrast between the so-called 'weak' /r/, generally a tap or approximant, and its 'strong' counterpart, /R/, generally a back fricative, spanning the entire range from velar to glottal. Here we follow the widely accepted convention of distinguishing them through upper and lowercase [11].

It is also useful to note that, although there is consensus as to the phonemic nature of BP stress [12], its relational, suprasegmental nature naturally excludes it from the co-occurrence counts.

4.2.3 Frequency counts

For each corpus, co-occurrence counts of phoneme pairs generated the following two 26 x 26 square matrices:

- Type count: performed on the word lists.
- Token count: performed on the entire corpora.

4.3. Statistics

4.3.1 Distance measure

To derive the distance measure, we first calculated the Spearman rank order correlation coefficient for all the variables in each co-occurrence matrix. From the correlation matrices, we derived the distance matrices by calculating 1-R. This measure assumes that distance is inversely proportional to correlation.

4.3.2 Multivariate exploratory techniques

Both of the techniques adopted are graphical ways of exploring relationships in a multivariate data set and make no assumptions about normality:

- Cluster analysis used single linkage as the amalgamation rule, i.e., the nearest neighbour was the criterion for linking different clusters.
- Multidimensional scaling used stress as the goodness of fit measure. No more than two dimensions were necessary for interpretable results to emerge.

The resulting graphs group segments with correlated co-occurrence frequencies. As predicted, such groups also share sonority-related properties.

It is important to note that, being means of exploring underlying relationships in quantitative data, neither technique tests for significance of the groupings.

5. RESULTS

For each statistical technique, we will first compare types and tokens for LAEL, and then for CETEN.

Some adjustments in the phonetic script are necessary, due to the inability of the statistical software to read IPA. Thus, we will use the Speech Assessment Methods Phonetic Alphabet (SAMPA) in the Figures. Here is the converted inventory:

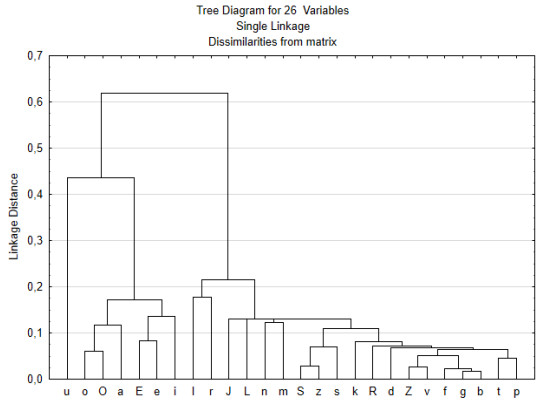
- Consonants: p, b, f, v, m, t, d, s, z, n, l, r, S=ʃ, Z=ʒ, J =ɲ, L=ʎ, k, g, R.
- Vowels: i, e, E=ɛ, a, O=ɔ, o, u.

5.1. Cluster analysis for LAEL

5.1.1. Types

Figure 1 shows the groupings for types. Note that Cs and Vs, as well as sonorants and obstruents, clearly split. The apparent exception, strong /R/, is actually a fricative. There is also a sharp split between front and back vowels. In addition, some remarkable consonant groups emerge: /l, r/, /J=ɲ, L=ʎ/, /n, m/, and /s, z, S=ʃ/.

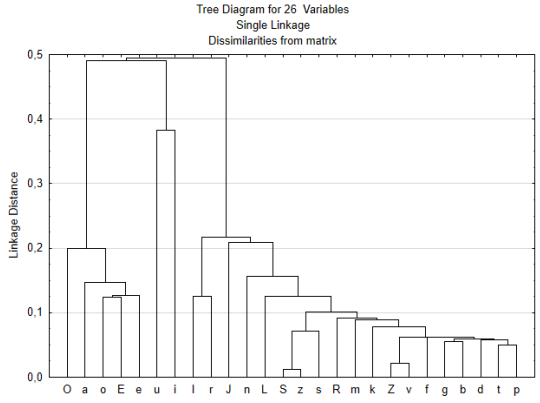
Figure 1: Clusters with single linkage for LAEL types.



5.1.2. Tokens

In Figure 2, tokens behave similarly, except for nasals, as /m/ moves right, linking with obstruents.

Figure 2: Clusters with single linkage for LAEL tokens.

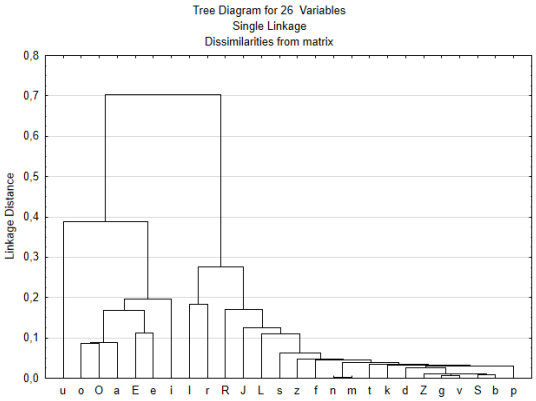


5.2. Cluster analysis for CETEN

5.2.1. Types

Figure 3 reproduces most of the LAEL patterns, except for the reversed positions of /R/ and /m, n/. A curious, isolated finding is the contiguity of /b, p/.

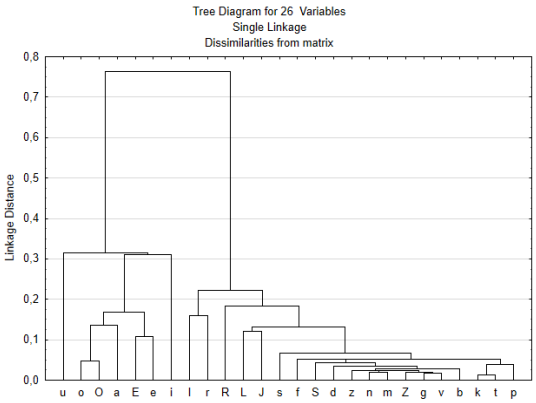
Figure 3: Clusters with single linkage for CETEN types.



5.2.2. Tokens

Similar patterns apply to tokens in Figure 4.

Figure 4: Clusters with single linkage for CETEN tokens.



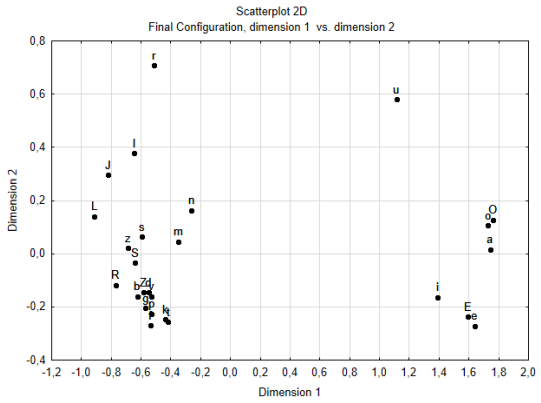
5.3. Multidimensional scaling for LAEL

5.3.1. Types

Figure 5 shows the two-D diagram for LAEL types. The first dimension is straightforward, as it separates vowels from consonants. The second dimension is trickier; as it seems to refer to sonority on the C side (left) and backness on the V side (right).

Patterns are similar to those observed in clusters. Note the cramming of obstruents in the 2-D space, which parallels their chain amalgamation in clusters, indicating consistency across techniques. This makes it clear that co-occurrence frequencies do not capture fine distinctions within the obstruent group.

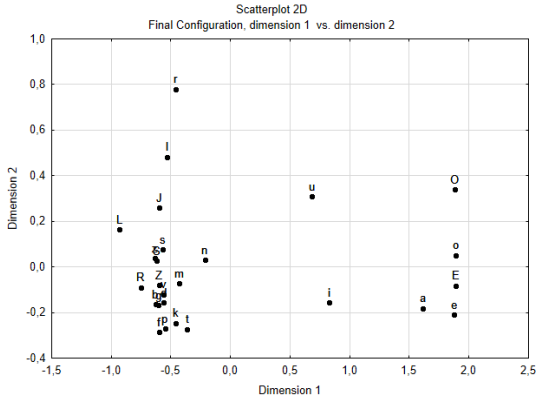
Figure 5: Two-dimensional space for LAEL types.



5.3.2. Tokens

Curiously, Figure 6 and 7 group the voiceless stops /p, t, k/ in the same way as the cluster of Figure 4.

Figure 6: Two-dimensional space for LAEL tokens.

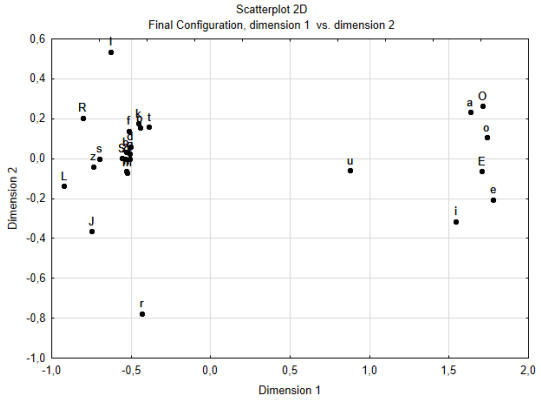


5.4. Multidimensional scaling for CETEN

5.4.1. Types

In Figure 7, patterns remain stable, even if obstruents are slightly more crammed.

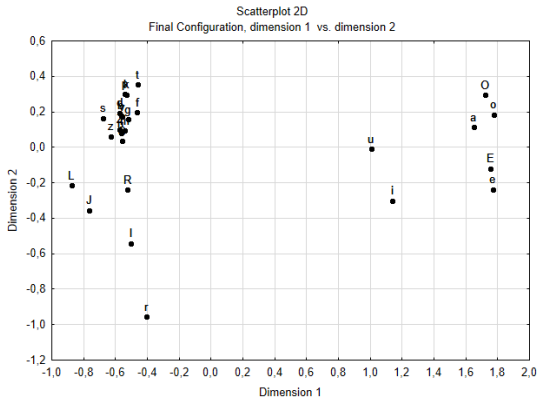
Figure 7: Two-dimensional space for CETEN types.



5.4.2. Tokens

It is worth noting that the vowels /i, u/ begin to get closer in Figures 6 and 7, and come their closest in Figure 8. This may reflect differences in sample size.

Figure 8: Two-dimensional space for CETEN tokens.



6. DISCUSSION AND CONCLUSION

The results support the initial hypotheses. Co-occurrence frequencies indeed seem to convey sonority distinctions between phoneme classes.

Consistency across counts, corpora and statistical techniques indicates that BP phoneme co-occurrence frequencies are stable, and, therefore, may predict groupings within the major classes. This may actually constitute an important bootstrapping mechanism for such classes, which participate in such important BP phonological processes as rhotacization of /l/ and vocalization of coda sonorants.

Although failing to distinguish obstruent subclasses, the patterns emerging in this study are highly structured and meaningful for sonorants. Recall that subclasses such as 'front vowels' or 'liquids' emerge because the co-occurrence frequencies of their members are highly correlated. Moreover, there is no further coding of the data. This suggests that phonotactics is sensitive to speech production constraints and thus differs from symbol combinatorics.

As clearly stated in the introduction, this study is only preliminary, inasmuch as the statistical techniques employed are simple visual ways of organizing the quantitative data and cannot test hypotheses about differences between the resulting phoneme groups. Our aim here was primarily to show that this is a fertile avenue for future research.

Further work will have to apply other, more powerful statistical classifiers to BP and other languages that follow the sonority principle. At the same time, it should look at languages that violate it.

7. REFERENCES

[1] Lisker, L. 1985. The pursuit of invariance in speech signals. *J. Acoust. Soc. Am*. 77, 1199–1202.

[2] Perkell, J., Klatt, D. (eds.), 1986. *Invariance and Variability in Speech Processes.* Hillsdale: Erlbaum Associates.

[3] Stevens K., Keyser S. 1989. Primary features and their enhancement in consonants. *Language* 65, pp. 81-106

[4] Lehnert-LeHouillier, H. 2009. Phonetic and phonological aspects of liquid devoicing in Thai, Hungarian, and American English stop-liquid sequences. *University of Rochester Working Papers in the Language Sciences* 5, 49-68.

[5] Strycharczuk, P. 2012. *Phonetics-phonology interactions in pre-sonorant voicing*. Unpublished doctoral dissertation. University of Manchester.

[6] Jusczyk, P., Luce, P., Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *J. of Memory and Language* 33, 630-645.

[7] Adriaans, F. 2011. *The induction of phonotactics for speech segmentation: converging evidence from computational and human learners.* Unpublished doctoral dissertation, Utrecht Institute of Linguistics.

[8] Almeida, A. 1976. The Portuguese nasal vowels: phonetics and phonemics. In J. Schmidt-Radefelt, (ed.), *Readings in Portuguese Linguistics*. The Hague: North Holland, 349-396.

[9] Pontifical Catholic University of São Paulo, Graduate Program in Applied Linguistics, http://www2.lael.pucsp.br/corpora/.

[10] Interinstitutional Centre for Computational Linguistics, University of São Paulo, São Carlos, http://www.linguateca.pt/cetenfolha/index_info.html.

[11] Cruz-Ferreira, M. 1995. European Portuguese. *J. Intern. Phon. Assoc.* 25, 90-94.

[12] Câmara Jr., J. 1972. *The Portuguese Language*. Chicago: University of Chicago Press.