SYLLABIC REDUCTION IN ITALIAN CONNECTED SPEECH: TOWARDS THE INTEGRATION OF LINGUISTIC AND COMPUTATIONAL APPROACHES

Loredana Schettino, Vincenzo Norman Vitale, Francesco Cutugno

University of Naples Federico II
{loredana.schettino, vincenzonorman.vitale, cutugno}@unina.it

ABSTRACT

In spontaneous speech, words are normally uttered in continuous chains whereby boundaries are blurred and segments can undergo changes, e.g., vowel centralization, consonant lenition, or deletion, due to structural linguistic factors or contingent sociolinguistic factors. However, such "reduction" processes are often studied by comparing connected speech realizations with isolated words in their expected, phonological forms. This study aims at investigating systematic patterns of phonetic variation in a dataset of spontaneous Italian discourse by considering the syllabifications of the speech chain at the phonological and phonetic levels. Moreover, since phonetic segmentation, based on human perception, implies a discretization of continuous productions, we employ unsupervised computational techniques, namely clustering algorithms, which allow for the observation of patterns emerging from the data without external (human) interference. The linguistic analysis shows systematic reduction processes related to the syllabic structure and lexical stress. The computational representation highlights a convergence of syllables to the non-marked structure in Italian (CV).

Keywords: reduction, syllabic structures, connected speech, clustering, unsupervised learning

1. INTRODUCTION

In spontaneous speech, the phonetic realization of words may vary with respect to their "canonical form" and speakers' utterances can be represented as continuous chains of words which, for the economy of production efforts, are commonly under-specified in the speech signal. This normally results in acoustic "reduction" which can manifest itself as spectral reduction, temporal reduction, and absence or change of segments [1] and ranges from subtle weakening up to segmental changes, i.e. from vowel centralization or consonant lenition to deletion of segments or even of multiple syllables [2, 3]. Speech reduction phenomena have been described in many different languages (a.o. German [4], Dutch [5], American English [6], Italian [7], French [2]) as representing, in fact, the rule rather than the exception in spoken language communication [8].

The type or extent of this variation may depend on sociolinguistic factors related to the communicative situation and the speakers [9] and linguistic factors, like prosodic features, lexical category, the discursive function [4, 10, 5]. In particular, a study on Italian [11] showed that vowel centralization represents a structural feature independent of diaphasic and diatopic variation. As for the listeners' processing of pronunciation variants, it was highlighted that in speech understanding, the frequency of occurrence of the variants with relation to specific contexts plays a more important role than the degree of reduction [3, 12]. Therefore, in-depth investigations of the patterns of phonetic variation in speech are crucial to deepen our understanding of speech production and comprehension mechanisms.

Reduction processes have often been studied by comparing, segment by segment, phonetic realizations with the phonological forms of isolated words, even when they appeared in connected speech. At the same time, evidence has been provided that the syllable can represent a relevant basic unit of speech production and perception rather than phonetic segments [13, 8]. In particular, [14] shows that the observable variation in connected speech is more systematic at the syllabic level than at the segmental one. Consequently, comparing the syllable sequence of the speech chain as expected in the phonological stream to that observed phonetically seems to be most suitable to describe reduction phenomena in connected speech.

In the last couple of decades, technological tools have been introduced to support the analysis of acoustic reduction by providing automatic and internally coherent phonetic annotations [2, 15, 1]. Further computational tools could be employed in the analysis of systematic phenomena in phonetic realizations. In particular, clustering algorithms are unsupervised tools that could be exploited to observe how segments identified as phonetic syllables will be grouped on the basis of the features extracted from the corresponding signal. In a recent work on the explainability of Deep Learning based Automatic Speech Recognition (ASR) systems [16], authors employed the K-Means algorithm to evaluate the similarity of information encoded in Deep Neural Networks (DNN) layers. However, the considered approach employs unknown features which are automatically extracted and encoded by DNN. Conversely, we aim to conduct an informed analysis by extracting specific features generally used in speech analyses [17, 18] through unsupervised computational tools in order to provide further support to the analysis of the patterns of phonetic variation in speech.

The main purpose of this study is to build upon the research proposed in [11] by investigating the phonetic variation patterns that may be observed in the speech chain, and their relation to specific linguistic structures, namely, syllabic structures and stress with the support of computational methods.

2. METHOD

2.1. Data and annotation

The investigation concerns the Italian dataset of the Nocando corpus [19] which consists of spoken narrative texts by 11 subjects (university students).

To obtain reliable annotations, the audio files and transcriptions were processed using the WebMAUS Basic services [20] and the provided phonological and phonetic transcriptions were manually edited in Praat [21] and syllabified according to the principles of sonority sequencing and onset maximization [22]. These principles were applied to both the phonological and phonetic representation of the speech chain, which means that

- resyllabification is expected between words when the second one begins with a vowel (e.g. *un'orecchio* [u.no.ˈre.kkjo];
- long consonants and geminates are considered as *onset* of the same syllable;
- the initial [s] of a consonantal nexus is assigned to the *coda* of the preceding syllable (e.g. *questo* [ˈkwes.to] in Italian). Exceptionally, it is considered *onset* only in the rare cases when it occurs at the absolute beginning of utterance.

So "phonological syllables" are those expected after resyllabification but before reduction phenomena, whereas "phonetic syllables" are those that are actually realised (see Fig. 1).

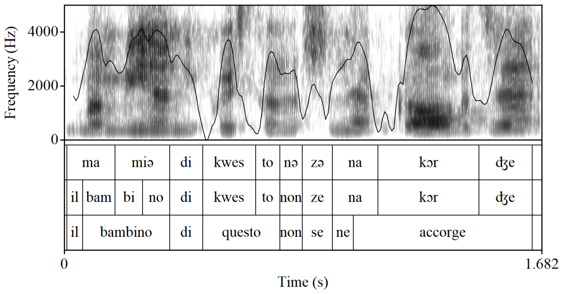

Figure 1: Example of annotation. I: phonetic level; II: phonological level; III: ortographic level.

2.2. Linguistic analysis

The alignment between the sequence of phonological syllables and one of the phonetic syllables was evaluated using SCLITE, which is a tool included in the Speech Recognition Scoring Toolkit (SCTK) provided by the American National Institute of Standards and Technology (NIST). The evaluation output reports the following cases:

- "Deletion", a phonetic syllable is found in place of two phonological ones;
- "Substitutions", a phonetic syllable is different from the corresponding phonological one;
- "Correct", the phonetic and the phonological syllables are alike.

The phonological syllables were also annotated for their phonetic structure (CV, CCV, CCVC, CVC, VC, V) and contextual saliency (lexical stress). For syllables subjected to substitution, the position of the syllabic structure in which it occurs is indicated, namely, Onset (O), Nucleous (N) and/or Coda (C).

To evaluate the role of lexical stress and syllabic structure on phonetic variation and control for individual variability, a statistical analysis was conducted on R [23] and Generalized Linear Mixed Models were fitted [24]. First, the evaluation levels were considered as binomial dependent variables; Syllabic Structure and Lexical Stress as interacting independent variables, and Speaker as a random effect. Then, on the subset of syllables subjected to substitution, the positions of the changed segments were considered as binomial dependent variables, Lexical Stress as an independent variable.

2.3. Computational Analysis

Considering the phonetic syllable as the base unit, we also assume the syllable is a data point represented with a discrete set of features. Based on the above-mentioned discrete representation, we employ a class of unsupervised learning approaches named *Clustering*. Approaches of this kind are meant to group similar items based on similarity criteria, which in our case means comparing syllables on the basis of a discrete set of features rather than the whole signal. The following clustering techniques, implemented in the Scikit-Learn toolkit [25], were considered:

- **K-Means** [26] is a vector-quantization method that divides n objects in k clusters based on their mean distance;
- **Hierarchical Agglomerative Clustering (HAC)** [27] is a greedy technique that aims at grouping (or splitting) clusters based on a similarity measure. The final output is a clusters hierarchy which could be divided based on the number of desired clusters.

These techniques require, in some way, an apriori expert intervention to determine the number of desired clusters. In our case, for both **HAC** and **K-Means**, the desired *number of clusters (k)* was fixed to the five most frequent classes of syllable structures among those found in the dataset described in Section 2.1, namely those with more than ten samples. Then, through the OpenSmile toolkit [17], we extract, for each considered sample, the GeMAPS set [18], composed of 62 features.

3. RESULTS

3.1. Linguistic Analysis

The dataset considered for the analysis consists of 940 syllables at the phonological level. Most of these are phonetically realized as expected (C=67%), a quarter of them have been subjected to variation (24%) and the rest have been deleted (D=9%). As expected, most syllables are unstressed (71%) and the stressed ones are fewer (29%). As for the syllabic structure, the following patterns are observed: CV (65%), CVC (16%), CCV (9%), VC (3%), CCVC (3%), V (2%).

The presence of lexical stress and the syllabic structure are both found to be significant predictors of phonetic variation (see Fig. 2 and 3). In particular, lexical stress predicts the occurrence of deletions, with unstressed syllables being most likely to be deleted (*Est*=-1.56, *SE*=0.40, *z*=-3.88, *p*=0.0001), but not the occurrence of substitutions (*Est*=-0.33, *SE*=0.18, *z*=-1.91, *p*=0.055). As for syllable structures, CV represents the most stable structure, which is significantly less subjected to variation (*Est*=-0.63, *SE*=0.28, *z*=-2.26, *p*=0.023), whereas V and VC structures are significantly more subjected to deletion (V: *Est*=2.43, *SE*=0.92, *z*=2.64, *p*=0.008; VC: *Est*=1.68, *SE*=0.81, *z*=2.06, *p*=0.039).

As for the cases of substitution, most changes concern either the syllabic onset (46%) or the nucleus (45%). Moreover, lexical stress (in Fig. 4) is found to predict which part of the syllable is affected by variation. Namely, the nuclei in unstressed syllables (*Est*=-1.13, *SE*=0.40, *z*=-2.82, *p*=0.005), whereas the syllable onsets in stressed ones (*Est*=0.81, *SE*=0.32, *z*=2.45, *p*=0.012).

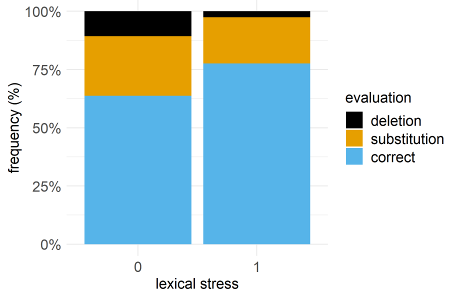

Figure 2: Frequency (%) of the evaluation output cases per stressed (1) and unstressed (0) syllables.

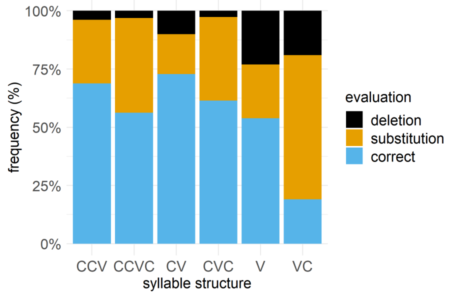

Figure 3: Frequency (%) of the evaluation output cases per syllabic structure.

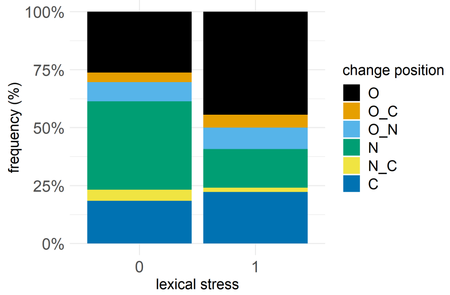

Figure 4: Frequency (%) of change position in the syllabic structure per lexical stress condition.

3.2. Computational Analysis

In table 1, we report results for each considered clustering technique. Each pattern label is associated with a cluster based on frequency; for example, the CV pattern label is assigned to the cluster with the most significant number of samples belonging to the CV pattern. Results highlight a particularly high similarity between the CV pattern, which is also the most frequent in the considered dataset, and the CCV and CCVC patterns (the K-Means assigns the CV label to most items per each class of syllabic structure, see the CV column in Table 1). This means that, as for the considered statistical features, the CV pattern is acknowledged to be very similar to the other two ones when considering algorithms requiring some prior knowledge about data, like the number of clusters for K-Means and HAC. Please note that behind all patterns a lot of acoustic information is encapsulated in both the onset *C* and *V*, which are representative symbols for different, respectively, consonant phones and vowel phones.

Algorithm	**Class**	CCV	CCVC	CV	CVC	VC
K-Means	CCV	**12**	6	21	17	9
K-Means	CCVC	0	**4**	9	6	0
K-Means	CV	85	128	**227**	55	81
K-Means	CVC	0	15	68	**35**	3
K-Means	VC	1	0	5	2	**1**
HAC	CCV	**21**	17	6	17	4
HAC	CCVC	0	**5**	4	7	3
HAC	CV	166	140	**174**	84	12
HAC	CVC	3	44	21	**49**	4
HAC	VC	2	2	3	2	**0**

Table 1: Clustering results comparison.

4. DISCUSSION

The comparison between the syllabic annotation of the speech chain at the phonological and phonetic levels provided insight into the reduction processes, or rather the systematic phonetic variation, occurring in connected speech net of the expected syllabic restructuring phenomena across word boundaries. The analysis of the systematic relation between the variation phenomena and the syllabic structural characteristics, that is, in this case, the presence of lexical stress and structural configuration in terms of vowel and consonant(s) patterns allowed for the description of reduction processes related to linguistic structures which are, to a certain extent, independent from sociolinguistic factors. As expected [11], unstressed syllables are most likely to undergo reduction phenomena. However, distinguishing between deletions and substitutions, we found that unstressed syllables can be deleted rather than stressed ones, but both, stressed and unstressed syllables, can be almost equally subjected to substitution. Then, further investigation on substitution phenomena highlighted that systematic differences concern the position of the change, in that the presence of lexical stress prevents vowel deletion or change (such as centralization, which concerns unstressed syllables) but allows for onset changes (such as lenition or assimilation). As for the syllabic structure, CV is confirmed as the most frequent structure [28] and the most stable and resistant to variation. Instead, V and VC structures are most prone to deletion and to restructuring processes in the speech chain. This seems also in line with Greenberg's observation that syllable onsets are generally preserved while coda or nuclear constituents are more frequently subjected to underspecification [14].

The computational representation supported this finding by highlighting that most syllables, even when belonging to different structural classes according to the manual annotation, are associated with the CV label, which means that more complex structures (CCV, CCVC, CVC patterns) show, in a way, a considerable degree of similarity with CV structures. In particular, the stability of the syllable onset seems to ensure the recognition of a consonantal onset. As for syllables with CCV and CCVC structures, probably, the system recognizes consonantal acoustic features but, since it doesn't account dynamically for internal variations, it assigns a generic C to complex onsets. As for CVC structure, in Italian, closed syllables are most likely to contain a sonorant coda, so the cluster could assimilate the whole sonorant group to a nucleus and **CVC** is associated with **CV**. This suggests that the acoustic realization of syllables tends to converge to more simple structures. Nonetheless, caution is needed when interpreting this result given the risk of oversimplification by a computational representation that relies on features depicting each syllable as a whole and not allowing for investigation of the internal structure. Follow-up studies will concern the in-depth analyses of different clustering processes. However, this study supports the importance of an informed investigation that relies on domain knowledge rather than purely statistical computations.

The presented findings result from the constructive integration between more traditional linguistic and computational approaches to the description of systematic phonetic variation in speech, which can be relevant not only to unveiling the functioning of speech production and comprehension mechanisms but also to improving the performances of speech technologies such as ASR systems. [1]

[1] This work results from the collaboration between the authors: Loredana Schettino, §§ 1, 2.1, 2.2, 3.1, 4; Vincenzo Norman Vitale, §§ 1, 2.3, 3.2, 4; Francesco Cutugno, §§ 1, 4.

5. REFERENCES

[1] B. Schuppler, “Automatic analysis of acoustic reduction in spontaneous speech,” Ph.D. dissertation, Radboud University Nijmegen, The Netherlands, 2011.

[2] M. Adda-Decker, P. B. de Mareüil, G. Adda, and L. Lamel, “Investigating syllabic structures and their variation in spontaneous french,” *Speech communication*, vol. 46, no. 2, pp. 119–139, 2005.

[3] M. Ernestus and N. Warner, “An introduction to reduced pronunciation variants,” *Journal of Phonetics*, vol. 39, no. 3, pp. 253–260, 2011.

[4] K. J. Kohler, “Segmental reduction in connected speech in German: Phonological facts and phonetic explanations,” in *Speech production and speech modelling*. Dordrecht: Kluwer, 1990, pp. 69–92.

[5] M. T. C. Ernestus, *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface*. Utrecht: LOT, 2000.

[6] K. Johnson, “Massive reduction in conversational American English,” in *Spontaneous speech: Data and analysis. Proc. 10th international symposium*, 2004, pp. 29–54.

[7] R. Savy, “Riduzioni foniche nel parlato. il ruolo della morfologia flessiva nellâinterpretazione del messaggio e nella comunicazione,” Ph.D. dissertation, University of Roma Tre, Italy, 1999.

[8] F. Cangemi and O. Niebuhr, “Rethinking reduction and canonical forms,” in *Rethinking reduction*, F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, and M. Zellers, Eds. Berlin: De Gruyter, 2018, pp. 277–302.

[9] M. Ernestus, I. Hanique, and E. Verboom, “The effect of speech situation on the occurrence of reduced word pronunciation variants,” *Journal of Phonetics*, vol. 48, pp. 60–75, 2015.

[10] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, “Effects of disfluencies, predictability, and utterance position on word form variation in English conversation,” *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1001–1024, 2003.

[11] R. Savy and F. Cutugno, “Hypospeech, vowel reduction, centralization: how do they interact in diaphasic variations,” in *Proc. XVIth International Congress of Linguists*, Paris, 1998.

[12] S. Brand and M. Ernestus, “Understanding reduced words. The relevance of reduction degree and frequency of occurrence,” in *Proc. 19th International Congress of Phonetic Sciences, Melbourne*, 2019.

[13] F. Albano Leoni, “The boundaries of the syllable,” in *The Notion of Syllable across History, Theories and Analysis*, D. Russo, Ed. Cambridge: Cambridge Scholars Publishing, 2016.

[14] S. Greenberg, “Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation,” *Speech Communication*, vol. 29, no. 2-4, pp. 159–176, 1999.

[15] M. Adda-Decker and N. D. Snoeren, “Quantifying temporal speech reduction in French using forced speech alignment,” *Journal of Phonetics*, vol. 39, no. 3, pp. 261–270, 2011.

[16] A. Prasad and P. Jyothi, “How accents confound: Probing for accent information in end-to-end speech recognition systems,” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3739–3753.

[17] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proc. 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[19] L. Brunetti, S. Bott, J. Costa, and E. Vallduví, “A multilingual annotated corpus for the study of information structure,” in *Grammar & corpora 2009*, 2011.

[20] T. Kisler, U. Reichel, and F. Schiel, “Multilingual processing of speech via web services,” *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[21] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [computer program].” *Online: http://www.praat.org*, 1999-2022.

[22] M. Nespor, *Fonologia*. Bologna: Il Mulino, 1993.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2022. [Online]. Available: https://www.R-project.org/

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] J. A. Hartigan and M. A. Wong, “Algorithm as 136: A k-means clustering algorithm,” *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[27] F. Murtagh, “A survey of recent advances in hierarchical clustering algorithms,” *The computer journal*, vol. 26, no. 4, pp. 354–359, 1983.

[28] S. La Torre, “Le sillabe del lessico italiano: tipi, pattern e disposizioni. saggio di interrogazione del dbsli (data base delle sillabe del lessico italiano),” Ph.D. dissertation, University of Chieti-Pescara “G. D’Annunzio”, Italy, 2005.