Machine Learning of Probabilistic Phonological Pronunciation Rules from the Italian CLIPS Corpus

Florian Schiel[1]*, Mary Stevens*[1]*, Uwe Reichel*[1]*, Francesco Cutugno*[2]

[1]Bavarian Archive for Speech Signals, Munich, Germany
[2]Università Degli Studi di Napoli Federico II, Naples, Italy
[1]schiel|mes|reichelu@phonetik.uni-muenchen.de, [2]cutugno@unina.it

Abstract

A blending of phonological concepts and technical analysis is proposed to yield a better modeling and understanding of phonological processes. Based on the manual segmentation and labeling of the Italian CLIPS corpus we automatically derive a probabilistic set of phonological pronunciation rules: a new alignment technique is used to map the phonological form of spontaneous sentences onto the phonetic surface form. A machine-learning algorithm then calculates a set of phonological replacement rules together with their conditional probabilities. A critical analysis of the resulting probabilistic rule set is presented and discussed with regard to regional Italian accents. The rule set presented here is also applied in the newly published web-service WebMAUS that allows a user to segment and phonetically label Italian speech via a simple web-interface.
Index Terms: Italian, CLIPS, pronunciation, machine-learning, dialect, MAUS

1. Introduction

Linguistic phonetic analysis typically seeks to construct models of speech production based on the manual collection and inspection of empirical speech data. Such models often consist of abstract processes leading from pure meaning to the physical speech signal. These processes are often of algorithmic (e.g. rule based) and sometimes of probabilistic nature (e.g. Optimality Theory).

In speech technology, on the other hand, the speech signal is essentially treated as the output of a complex statistical source. The task is to model this source given certain (discrete) concepts (often called "training") and to use the resulting models to calculate conditional probabilities for different speech signals with different concept values ("test", e.g. [16], [13]).

In this sense, while there are exceptions on both sides, linguistic models tend to be "hypothesis-driven" while speech technology can be seen as "data-driven".

Both approaches also have their advantages and disadvantages. Linguistic models not only predict speech surface forms but allow the researcher to hypothesize about the nature of the production process, while technological models often neglect human speech production entirely. Technological speech models are *per se* adaptable to speakers, different situations or even new languages, whereas linguistic models are often language-dependent or based on the observation of small groups or even single speakers. Linguistic models often tend to be under-specified in the sense that very unlikely or even impossible surface forms are predicted with the same probability as more realistic ones (for instance recursive grammar rules), while technological models do the opposite: they adapt closely to the training material and might therefore not be robust enough to serve as a general model for speech (e.g. when applied to speech from a different domain).

In this contribution we suggest a blending of technological and linguistic approaches. We exemplify our approach taking the concrete problem of the pronunciation of contemporary Italian. Both paradigms outlined above must be able to relate linguistic (here: phonological) structures to spontaneous Italian words and sentences. For instance the word *praticamente* 'practically' can be transformed into the phonological form /pratikamente/[1] according to more or less agreed upon (standardized) pronunciation rules. But looking at real world recordings of Italian speech we find phonetic realizations such as [badigaente] or [priamente] and many others[2]. A linguistic model must explain this variability in pronunciation, and the classical approach would be to formulate phoneme-based replacement rules and tie their application to contexts e.g. "word-position", "speech rate", "dialect" (e.g. for [badigaente] there is voicing of intervocalic voiceless stops (/p, t, k/ > [b, d, g]/V_V) in many centro-southern varieties and deletion of intervocalic nasals (m, n > ø/V_V) in fast or less careful speech).

A technological approach on the other hand would use a statistical model (e.g. constraint-based Hidden Markov Models for words) to predict the probability of variation in pronunciation, based on the degree of variability observed in the training material.

In the following we show that both approaches can be (partly) merged. The basic idea is to use the phonological framework of the classical replacement rule together with machine learning techniques derived from speech technology to end up with a combined phonological-probabilistic model for Italian pronunciation. In taking this approach we aim to provide some new insights for linguistic research as well as to improve the performance of technological applications such as automatic speech recognition or automatic phonetic segmentation.

2. Data: The CLIPS Corpus

CLIPS (Corpora e Lessici dell'Italiano Parlato e Scritto) is a corpus of spoken Italian covering a broad variety of contemporary Italian speech ([12])[3].

To guarantee that the five main variants of Italian were rep-

[1]Here and in the remaining article all phonological and phonetic symbols are coded in the SAM Phonetic Alphabet [17].

[2]Examples taken from phonetic transcriptions of the MapTask recordings DGmtB03P and DGmtB04F, respectively, of the CLIPS corpus [12]; diacritics stripped from the original transcription.

[3]A complete description of all aspects of the project can be found in the website documentation (http://www.clips.unina.it).

resented the following cities, listed by relevant geo-linguistic area, were chosen as recording sites:
North: Turin, Genoa, Milan, Bergamo, Parma, Venice
Centre: Florence, Rome, Perugia
Upper southern: Naples, Bari
Extreme southern: Catanzaro, Lecce
Islands: Palermo, Cagliari
Informants were undergraduate students, aged between 18 and 30, who had always been resident in the relevant city area, and whose parents had also always been resident there. Males and females are, on average, equally represented in the corpus.

CLIPS is divided into the four sub-corpora: free-field recordings, radio recordings, television recordings, and telephone conversations. These parts vary along the diamesic (= "through the media") axis covering a wide range of possible recording channels (see Table 1). In total CLIPS comprises about 100 hours of audio recordings, the free-field section (mainly formed by dialogues, as we will see further on) is the largest and forms approximately 50% of the entire corpus; remaining sections cover about 16 hours of recordings each, with the exception of the ortho-phonic corpus that consists of less than 4 hours.

In particular, the free-field sub-corpus consists of elicited and (semi-)spontaneous dialogues (presenting a low level of formality) and read speech (further subdivided into word lists and sentence lists). The dialogue sub-part contains two types of recordings, elicited using two different techniques: the "Map Task" (MT) method ([2], [5]) and the "Spot-the-difference Game" (SD).

Table 1: *Corpus stratification with regard to three dimensions*

Diaphasic/Diamesic	Diatopic	Content
Dialogue (elicited)	15 regional varieties	map-task spot the difference
Read Speech	15 regional varieties	read sentences word list
Radio & TV	15 regional varieties	broadcast talk show commercials culture
Telephone	15 regional varieties	Auto WoZ
Ortho-phonic	standard	read sentences

A group of about 15 phoneticians orthographically transcribed about 30% of the recorded material, while 30 dialogues from the free-field sub-corpus (MT) were also labeled and segmented phonetically. CLIPS provides different types of segmental labeling delivered in TIMIT ([18]) format each corresponding to a phonetic/phonological/lexical level of analysis. Segmental time-aligned labeling includes the following levels (from narrower to broader):

1. acoustic ACS;
2. phonetic PHN;
3. phonological STD (citation forms);
4. lexical WRD;
5. extra-text ADD (comments).

ACS contains time references and labels related to the phases of stop and release of the occlusive and affricate consonants. PHN contains a broad segmental phonetic (acoustic) transcription (level 4, [3]). STD is a labeling level without time references containing a word by word phonological transcription automatically generated by means of a rule-and-exception based algorithm specifically produced within the CLIPS project. WRD consists of a redundant word by word orthographic transcription including labels for breaks, disfluencies, noises, and other similar phenomena.

In the present work we use the PHN and STD label sets of 30 MT dialogues with 30 participants (2 speakers deliver two MT dialogues in each of 15 recording sites) resulting in a total of 3229 transcribed dialogue turns, 32255 words and 87057 phonetic segments. Diacritical information was stripped from the PHN tier before our analysis, e.g. the nasalization of a vowel was not considered here (see Section 5).

3. Phonological to Phonetic Aligner

The alignment between a canonic transcription v and a connected speech transcription w is derived from their Levenshtein distance, i.e. the minimum edit costs to transform v into w. Following the *PermA* approach of [11] edit costs c for the edit operations *substitution, deletion* and *insertion* are defined in terms of conditional probabilities reflecting phone co-occurrences between canonic and connected speech transcriptions (including empty phones _) as follows:

- **Substitution:**
$$c(v_i, w_j) = \begin{cases} 0 & : \text{equal}(v_i, w_j) \\ 1 - P(w_j|v_i) & : \text{else.} \end{cases}$$
- **Deletion:** $c(v_i, \_\_) = 1 - P(\_\_|v_i)$
- **Insertion:** $c(\_\_, w_j) = 1 - P(w_j|\_\_)$

The probability model for the cost function is calculated on a word bigram list of canonical and spontaneous speech transcription pairs, which is derived by moving a bigram window with step-size 1 along the parallel word-segmented transcription data. The choice of bigram units instead of uni-grams serves to capture phonological processes operative across words.

PermA smoothes the conditional probabilities by weighting the co-occurrence count increments within a triangular window of length 3 and area 1. Insertion and deletion operations are treated the same way as substitutions by introducing the empty symbol '_'. Respective co-occurrences are counted by '_'-symbol padding and permutation in the shorter sequences in the alignment training pairs, normalizing the count increments of each permutation instance by the number of permutations.

After the training of the cost function, the aligner is applied to the parallel utterance transcriptions of our data in order to produce an output of the following form:

```
# s i # s o p r a # a u n # d a d o # p a ss a #
# s i # s o b r _ # a _ n # d a d O # b a ss a #
```

(First line being the phonological pronunciation, second line the phonetic transcription; '#' denotes the word boundary anchors and '_' a missing element.)

4. Probabilistic Micro Rule (PMR) Learning

In a first pass we went over the aligned data set and segmented the stream of mapped phonemic symbols into "matched" and "non-matched" sequences. Each instance of "non-matched" is then formulated into a Probabilistic Micro Rule (PMR) of the

form $a, x, b \rightarrow y$ where x is the non-match sequence from the phonological stream, y the corresponding sequence from the phonetic stream, and a and b the pre- and post-context sequence (which must match in both streams) of a fixed length cl. For example: $we, s, t, i \rightarrow ss$. Both, x and y may be or contain the empty element '_' but not the word boundary element '#' (since these always match); a and b in turn may not contain the empty element '_' (since these never match). In this study the context length cl of a and b was set to 1, since the amount of transcribed data in the CLIPS corpus was not enough to derive statistics about larger contexts. For each found PMR the total number of instances $N(a, x, b \rightarrow y)$ in the corpus was counted.

In a second run over the same data we count instances of the left side of each stored PMR $N(a, x, b)$ in the phonological stream only. The conditional probability for a rule application is then calculated as:

$$P(a, y, b|a, x, b) = \frac{N(a, x, b \rightarrow y)}{N(a, x, b)} \quad (1)$$

PMRs together with their respective conditional probability and the absolute count $N(a, x, b \rightarrow y)$ are then stored into a PMR set.

Since the annotations as well as the automatic mapping contain errors, a considerable proportion of the learned PMRs do not represent "real" events but are simply caused by errors. It is reasonable to expect that these errors are usually not repeated resulting in exactly the same PMRs (although systematic errors could theoretically occur, if for instance one human labeler systematically confuses phonemic symbols). A simple pruning algorithm deletes all PMRs with an absolute count less than a pruning threshold $N(a, x, b \rightarrow y) < T$. The optimal pruning threshold T depends on the size of the analyzed corpus. For the CLIPS corpus we found $T = 4$ a reasonable value resulting in mostly sensible PMRs.

The above procedure, applied to the CLIPS data set as described in Section 2, results in a set of 588 PMRs; as an example we list the 8 PMRs with the highest conditional probabilities learned from the CLIPS corpus:

```
   a,n,g>a,N,g   0.74961
  ja,n,k>ja,N,k  0.73531
  SS,E,n>SS,e,n  0.70121
   #,S,i>#,SS,i  0.70121
  a,dZ,i>a,ddZ,i 0.62998
   k,o,d>k,O,d   0.59821
   #,o,m>#,m     0.48034
we,s,t,i>we,ss,i 0.46082
 o,z,E,g>o,s,e,g 0.46082
   u,n,g>u,N,g   0.43987
```

Note here that the conditional probability does not necessarily reflect the usefulness/importance of the learned PMR for the model; a PMR with a very low conditional probability can nevertheless be useful if the probability for the condition $P(a, x, b)$ is high. Despite this we observed that in general PMRs with lower conditional probabilities tended to model more and more unlikely replacements; therefore in the following phonological analysis we focused on the first 290 PMRs (approx. half of the set).

The complete PMR set thus derived from the CLIPS corpus can be downloaded with the MAUS package from the Bavarian Archive for Speech Signals.[4]

[4]http://www.bas.uni-muenchen.de/forschung/Bas/

5. Phonological Discussion

The CLIPS phonetic PHN labels were chosen from a predetermined set that included (along with all the Italian allophones) allophonic variants resulting from coarticulation, known regional variants, and "unintentional variants" (see [12] for detail). The PHN transcription also contains diacritics such as glottalization and nasalization, but as noted earlier we did not consider diacritics in this investigation. Note that the PMRs reflect a mismatch only between phonetic labels and phonological/canonical labels. In this Section we want tease apart the possible explanations e.g. coarticulation, post-lexical sandhi or speech errors. We investigated

1. to what extent the PMRs correspond to linguistic descriptions of spoken Italian, which are many, cf. [12] for references; we also refer in particular to [1, 4], and
2. whether the PMRs can be classified as regional vs. universal.

Contemporary standard Italian is spoken with distinct regional accents even in relatively formal contexts by most native speakers (e.g. [4, 12]), and the PMR set within the MAUS tool should ideally be robust enough to cope with this kind of phonetic variation.

290 of 588 learned PMRs, in order of descending conditional probability, were manually classified into the 12 categories shown in Table 2 (we classified PMRs where x and y contained more than one element e.g. `d,e,ll,a>d,E,l,a`, in terms of the change affecting the consonant). The 17 PMRs classified as Nasal assimilation (all reflecting a predictable process before homorganic stops in Italian) and Other (mostly involving vowel v. glide alternations) are excluded from this discussion. For each of the remaining 273 individual PMRs, we also listed the filenames in which they occurred, allowing us to examine regional distribution. Vowel deletions and vowel

Table 2: *Probabilistic micro-rule type statistics*

PMR type	count	frequency of application
V Height	89	1568
V Deletion	48	1321
C Doubling	43	988
C Deletion	29	425
C Voicing	16	331
C De-gemination	15	437
C Lenition	13	216
C Assimilation	10	134
C Devoicing	9	76
C Fortition	1	8
Nasal assimilation	11	444
Other	6	177
Total	290	6125

height changes are the most common PMRs in terms of absolute frequency of application and are common to all cities in the data set (i.e. universal). Most PMRs involving vowel (and glide e.g. `v,ai,#>v,a,#`) deletions occurred at word boundaries, as expected for Italian ([4]). Vowel height changes all involved the mid vowels /E e/ and /o O/; the wide regional distribution of these PMRs corresponds well with ([1]) which states the merge

of [e E] and [o O] in spontaneous speech is a non-regional feature. ([1] also identify assimilation of /rC/>[C:] as a typical non-regional feature of spoken Italian, but we found no evidence of this particular assimilation in the PMR set.)

Consonant doubling was frequent in the corpus, but 37 of these 43 PMRs involved word-initial position (e.g. #,b,a>#,bb,a) likely reflecting the post-lexical sandhi process *raddoppiamento sintattico* (RS) (e.g. [4] p. 135, [10], [9]) under which certain words trigger word-initial consonant lengthening e.g. *tre* [g:]*atti* 'three cats' (with RS) but *due* [g]*atti* 'two cats'. RS is not learnable for our model because PMRs are only sensitive to one preceding segment - in this case a word boundary. In other words the PMRs do not allow us to distinguish RS from other cases of non-canonical doubling.

Other than in word-initial position, doubling was infrequently transcribed (although it was included in the closed allophonic set during segmental labeling) and was rather evenly distributed across regions, which is unexpected given non-canonical gemination is a southern Italian feature.

Consonant deletions typically affected intervocalic /m n v/ and based on our data can be seen as features of standard Italian common to all regions (in line with [1]).

PMRs involving voicing of velar stops in word-initial position were particularly frequent for data recorded in the south (e.g. of the 45 occurrences of #,k,e>#,g,e, 15 were from Palermo and 12 were from Rome). To the best of our knowledge voicing of /k/ in utterance-initial position is not associated with a particular regional variety and in fact, as noted above, our PMRs only refer to the immediately preceding segment, i.e. the word boundary #. The velar stops in question are very likely in intervocalic position across the word boundary, and voicing of intervocalic /p t k/ is widespread in centro-southern Italy both within and across word boundaries (e.g. [7]).

PMRs classified as consonant de-gemination were also relatively evenly distributed across cities, the most frequent involving lateral /l/ in articles e.g. a,ll,a>a,l,a which occurred 130 times.

All but two of the PMRs classified as consonant lenitions involved de-affrication of /tS/ and /dZ/, which linguistic descriptions ascribe to speakers from Rome, Perugia and Florence ([12, 4]) but we find to be a more widespread tendency in spoken Italian (half the 202 occurrences of these PMRs occurred in speech recorded in other cities).

Similarly devoicing of /z/ (e.g. o,z,E,g>o,s,e,g) was not confined to Roman Italian, as [4], p. 33 report, and instead appears to be a more widespread tendency for speakers from the centre-south (and 4 of 76 cases were found in data from northern cities).

Eight PMRs classified as consonant assimilation involved [st]>[ss] (e.g. we,s,t,i>we,ss,i) and this rule was universal, consistent with [1].

PMRs corresponding to voicing of geminate /ts:/ (associated with the Milanese accent [4]) and affrication of /s/ after nasals (associated with Tuscan and Roman [4], listed as Fortition in Table 2), were found in the speech data but not frequently, and mostly outside of the cities in which we would expect them to occur.

A well-known regional pronunciation not reflected in the complete set of 588 PMRs is the *gorgia toscana* (e.g. [6]). Expected only for Florence this was likely not detected by our approach, because diacritics were deleted from the CLIPS transcripts before analysis.

Overall, the PMR set corresponds reasonably well with linguistic descriptions of spoken Italian connected speech processes, but their regional distribution is in some cases more widespread than we would expect based on the descriptive linguistic literature. This unexpected homogeneity in the data suggests that the MAUS tool for Italian should be robust enough to cope with variation due to regional accents. On the other hand, the distribution of the PMRs did not reliably distinguish recordings from individual cities or major linguistic iso-glosses (e.g. northern varieties with de-gemination vs. centro-southern varieties with geminates and a tendency towards non-canonical gemination).

6. Application in MAUS

The HMM-based MAUS tool allows either the simple alignment of a phonetic transcript to the speech signal or the usage of a phonological or statistical (PMR) rule set to create a hypothesis space of possible pronunciations, which is then searched by a standard Viterbi algorithm (for details about the integration of acoustic and PMR probabilities into a probabilistic graph see [15]). MAUS has been shown to perform within the capabilities of human labelers ([14]) for German, for which a PMR set has been learned from the German Kiel Korpus ([8]). MAUS has been subsequently adapted to a number of other languages (currently 11), but until recently only German and Australian English made use of probabilistic PMR sets.

The PMR set based on the Italian CLIPS corpus as described in Section 4 has been incorporated into the current version of MAUS. To improve its robustness in new domains including different dialects and speaker characteristics, the PMR set was further reduced to 128 rules (using a pruning threshold of 20 instead of 4), and the left/right contexts were permutated to phonetically similar contexts. This results in a final PMR set of 764 rules. The Italian MAUS HMM set (based on an extension of the official SAMPA Italian phoneme set) was also re-trained to the manual segmentations in the CLIPS corpus.

At the time of writing no independent benchmark set was available for Italian to formally evaluate the improvement of the MAUS labeling and segmentation for Italian speech. An informal application to read and spontaneous speech from the non-labeled part of the CLIPS corpus as well as to other Italian recordings performed at the BAS showed an improvement to both the phonetic transcript and the segmentation of individual phones. The Italian module of MAUS was recently incorporated in the new CLARIN service *WebMAUS* allowing the user to apply the MAUS technique to Italian speech recordings via a web interface, thus avoiding the need to install the MAUS software on a local computer. The web interface can be accessed via "clarin.phonetik.uni-muenchen.de/BasWebServices".

7. Conclusion

The proposed mixed phonological-technical approach to analyzing the pronunciation of spontaneous Italian yielded on the one hand interesting insights into the contemporary application of well-known phonological processes, and on the other hand a useful technological resource in the form of a set of machine-learned probabilistic pronunciation rules. Since the entire analysis was carried out automatically, the same analysis can be applied to other languages for which a sufficiently large corpus of spontaneous speech is available. Our analysis of phonological processes showed that some phonetic variants were identified outside the specific regions with which they are associated in the descriptive literature and in doing so sheds new light on the contemporary pronunciation of standard Italian.

8. References

[1] Albano Leoni F, Maturi P (1994): Didattica della fonetica e parlato spontaneo. In: Ramat A G, Vedovelli M (Eds.) 'Italiano: lingua seconda/lingua straniera', Pubblicazioni della Societa di Linguistica Italiana Vol 34. Rome, Bulzoni: 153-164.

[2] Anderson A H, Bader M, Bard E G, Boyle E, Doherty G, Garrod S, Isard St, Kowtko J, McAllister J, Miller J, Sotillo C, Thompson H, Weinert R (1991): The HCRC Map Task Corpus. Language and Speech, 34(4): 351-366.

[3] Barry W J, Fourcin A J (1992): Levels of Labelling. Computer Speech and Language, volume 6, issue 1, pp. 1-14.

[4] Bertinetto P M, Loporcaro M (2005): The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. Journal of the International Phonetic Association 35(2): 131-151.

[5] Carletta J, Isard A, Isard St, Kowtko J, Doherty-Sneddon G, Anderson A (1996): HCRC Dialogue Structure Coding Manual. Human Communication Research Centre, University of Edinburgh, Technical Report HCRC/TR-82.

[6] Dalcher C V (2008): Consonant weakening in Florentine Italian: a cross-disciplinary approach to gradient and variable sound change. Language Variation and Change, Vol 20: 275-316.

[7] Giannelli L, Cravens T D (1997): Consonantal weakening. In: Maiden M, Parry M (Eds.): The dialects of Italy. London/New York: Routledge. pp 32-40.

[8] Kohler K (1996): Labelled Data Bank of Spoken Standard German - The Kiel Corpus of Read/Spontaneous Speech. In: Proceedings of the ICSLP 1996, Oct 3-6, Philadelphia, PA, USA.

[9] Loporcaro M (1997): L'origine del raddoppiamento fonosintattico: saggio di fonologia diachronica romanza. Basel/Tuebingen: Francke Verlag.

[10] Stevens M (2012): A phonetic investigation into "Raddoppiamento Sintattico" in Sienese Italian Speech. Bern: Peter Lang.

[11] Reichel U (2012): PermA and Balloon: Tools for string alignment and text processing, Proc. Interspeech. Portland, Oregon, paper no. 346.

[12] Savy R, Cutugno F (2009): CLIPS Diatopic, diamesic and diaphasic variations in spoken Italian. 5th Corpus Linguistic Conference, Liverpool.

[13] Schiel F, Kipp A, Tillmann H G (1998): Statistical Modeling of Pronunciation: It's not the Model, it's the data; Proceedings of the ESCA Tutorial and Research Workshop on 'Modeling Pronunciation Variation for Automatic Speech Recognition', May 1998, Kerkrade/Netherlands.

[14] Schiel F (1999): Automatic Phonetic Transcription of Non-Prompted Speech, Proc. of the ICPhS 1999. San Francisco, August 1999. pp 607-610.

[15] Schiel F, Draxler Chr, Harrington J (2011): Phonemic Segmentation and Labeling using the MAUS Technique. Workshop 'New Tools and Methods for Very-Large-Scale Phonetics Research', University of Pennsylvania, January 28-31, 2011.

[16] Tajchman G, Jurafsky D, Fosler E (1995): Learning Phonological Rule Probabilities from Speech Corpora with Exploratory Computational Phonology. In Proceedings of ACL 95, Cambridge, MA, pp 9-15.

[17] Wells J C (1997): SAMPA computer readable phonetic alphabet. In: Gibbon D, Moore R, Winski R (eds.): Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.

[18] Zue V, Seneff S, Glass J (1989): Speech database development: TIMIT and beyond. In: Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, The Netherlands, pp. 35-40.