On the Evaluation of Marginal Improvements in Pronunciation Variation Modeling for Spanish

J. Macías-Guarasa, J. Ferreiros, R. San-Segundo, J.M. Montero, R. Córdoba and V. Sama

Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica
Universidad Politécnica de Madrid

{macias, jfl, lapiz, juancho, cordoba, vsama}@die.upm.es

Abstract

In the context of large vocabulary speech recognition systems it is of major importance to accurately model the allophonic variations to be faced in a real world task. Evaluation of which variants are actually improving the system performance is crucial, as it determines the acceptance of the pronunciation alternatives used.

Traditional approaches in this direction use different criteria and, typically, evaluation only cares about the global impact of the augmented dictionaries in the error rate, so that this lead to little further insight on till what extent the proposed variations are actually working or not. Our proposal in this paper is also evaluating the marginal improvement due to every pronunciation variation used (initially restricted to rule-based variant generation), defining specific improvement metrics.

We experimentally show how these metrics actually show the improvement achieved by the application of rules when dealing with certain pronunciations (or speakers in general) while their global impact in error rate may not be statistically significant.

1. Introduction

In the literature there are plenty of references to the problem of introducing multiple pronunciations (an excellent revision can be found in [1]). The general idea is facing the variability found in the produced speech signals due to cultural and dialectal differences and articulation modes specific to certain speakers.

Strictly speaking, almost every speech recognition system implicitly deals with pronunciation variations, given that the acoustic models used (typically HMMs) take into account all the segmental and temporal variations. However, what we really want to do is explicitly modelling the pronunciation alternatives, which has been proved to achieve substantial improvements if the acoustic models closely match the transcriptions [2]. The most usual strategies to add pronunciation variants use either knowledge based approaches, applying phonological rules to the canonical dictionaries, or data-derived pronunciation variants [1].

In spite of the intense work in this area, there are no definitive solutions yet. When you add pronunciation variations in the task lexicon (dictionary), the objective is improving the acoustic decoding process in the speech recogniser. However, if the added variants are not adequate, the final error rate can actually increase, as the dictionary is bigger. With this restriction, the research teams are extremely careful when adding variants and several approaches to generate and limit their number have been proposed, such as using a maximum likelihood criterion [4], smoothing the automatically derived phonetic transcriptions [5], measuring the occurrences of added variants [6] or using confusability metrics [3], to name a few. All of them try to evaluate (either implicitly or explicitly) till what extent the added variants are actually achieving better results or not.

In this paper we propose a novel approach for evaluating the (marginal) improvements every pronunciation variant can actually achieve, so that we could use such evaluation to decide which variants to select in the final system. Our proposal is also meant to allow us to do a more detailed error analysis, as it gives more insight on the efficiency of certain phonological variations.

2. Experimental setup

In this paper, we will only consider pronunciation alternatives at the segmental level, and, more specifically, within word variations. In our previous work, we have applied both knowledge-based and data-driven strategies, but in this paper we are only dealing with rule-based pronunciation variation generation, as it best serves our purposes.

This study will be done on an isolated word task, and this could be seen as a major drawback, as we are mainly interested in evaluating continuous speech recognition tasks. However, almost every study on pronunciation variation available in the literature, is based on generating phonological pronunciation variations at the word level (although there is also

recent work dealing with pronunciation variations at the lexical level [9]), so that the results obtained on an isolated word task should be extensible to continuous speech ones, specially given that we are using within word variations (our previous experience described in [11] showed very different results between the continuous speech and isolated word cases, but in that work the rules applied were explicitly designed to cope with between word coarticulation effects).

2.1. Speech recognition system

In order to have a reasonably broad experimental scenario, we decided to use a speech recognition system based on the hypothesis-verification paradigm. The hypothesis module follows a bottom-up approach and its architecture is shown in Figure 1. Its main modules are (a more detailed description can be found in [7]):

Acoustic Processing: The input speech signal is preprocessed obtaining a vector of parameters composed of 8 MFCCs, 8 delta-MFCCs, log-energy and its first derivate. They are soft quantized in order to use semi-continuous (SCHMMs), with 2 codebooks and 256 centroids each.

Phonetic String Build-Up: This generates a string of alphabet units using a frame-synchronous one-pass algorithm. The alphabet is composed of 45 allophone-like context-independent HMMs plus 2 additional silence units.

Lexical Access (LA): The phonetic string is matched against the whole dictionary, using a dynamic programming algorithm and trained alignment costs.

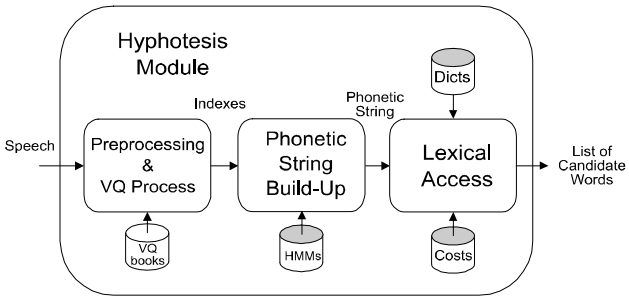

Figure 1: Hypothesis module schematic diagram

The verification module (not shown in figure 1), which receives a list of candidate words (sorted according to their likelihoods) generated by the hypothesis module and generates the final recognition result (which can be a single result or, again, a sorted list of words), is based on the viterbi algorithm, using context-dependent models.

This architecture has two main advantages concerning pronunciation variation related research:

1. The hypothesis module contains an automatic phonetic transcriptor, which can be used to generate data-driven pronunciation variants
2. The evaluation of phonological alternatives can be done in the hypothesis module (reasonably simple, with low computational demands and with low performance); and the verification module (achieving much lower error rates, but demanding higher resources). This allows us to experiment on a broader range of classifier conditions.

2.2. Databases and dictionaries

In our experiments, we have used part of the VESTEL database [8], a realistic speaker-independent speech corpus collected over commercial telephone lines, composed of digits, numbers, commands, city names, etc. organized in two sets:

- PRNOK5TR: Devoted to generic system training and composed of 5,820 files
- PERFDV: Devoted to system testing using the models generated with PRNOK5TR and composed of 2,536 utterances

We have also applied the leave-one-out method on a set composed of 9,790 utterances (the ones in PRNOK5TR, PERFDV and an additional subset of 1434 utterances), in order to increase the statistical significance of the results (10 subsets). We will refer to this set as LOO.

In the tasks described in this paper, we have used dictionaries composed of 1175 words for the PRNOK5TR and PERFDV sets and 1952 for the LOO one. The original system handles dictionaries up to 10,000 words, but in order to apply our proposal, we needed to restrict the dictionaries to words for which we actually had acoustic samples. We will refer to these dictionaries as the "canonical" ones.

2.3. Rule selection process

From internal studies carried out in our Group, we generated an exhaustive repertoire of Castilian Spanish variants, elaborating an ambitious collection of relatively general dialectal variations, in order to be efficiently included in speech recognition systems.

In our case, we further reduced this repertoire (as it generated too many variants with limited impact and importance), in line with the ideas shown in [10] and with a clearly pragmatic view, leading to a selection of twelve phonological rules, shown in Table 1.

Additional experiments selected a rules subset, composed of five rules (marked in bold in Table 1) as the most relevant ones. The selection criteria considered the relative increment in dictionary size:

- Rremoving all rules with negligible or null impact (*bs, equis, Gn, kt, pt, kT*).
- Removing the *ceceo* rule due to the high number of additional pronunciations introduced.

In the following sections we will refer to this subset as *"selection of rules"*.

3. Evaluation strategy

Traditionally, the evaluation of the introduction of pronunciation variations is based on measuring the global impact in word error rate (WER). This typical approach is somehow limited as it does not allow an easy comparison between systems [1]) neither it gives any insight on the actual impact of every single rule on the system performance, unless detailed error analysis is carried out (which is often omitted due to a number of practical reasons [1]). To overcome these limitations, in this work we propose a whole set of metrics that complement the simple WER evaluation and allow us to perform detailed error analysis.

Table 1: Phonological rules studied.

Rule name	Description
bs	[bs] → [s]
ceceo	[θ] → [s]
seseo	**[s] ® [q]**
dfinal	**final [d] deletion ; final [d] ® [q]** **final [d] ® [t]**
participio	**intervocalic [d] deletion in participles**
equis	[ks] → [s] ; [ks] → [gs]
Gn	[Gn] → [Xn] ; [Gn] → [n]
hue	**[we] ® [Gwe]**
kt	[kt] → [Xt] ; [kt] → [gt] ; [kt] → [Tt] ; [kt] → [t]
pt	[pt] → [bt] ; [pt] → [Xt] ; [pt] → [gt] ; [pt] → [t]
kT	[kT] → [XT] ; [kT] → [gT] ; [kT] → [T]
sfinal	**final [s] deletion**

3.1. Evaluation proposal

In our approach, the evaluation should consider:

- The relative variation in error rate, measured on the whole test set
- The effective marginal improvement of the considered variant, achieved on the subset of the whole test set which could benefit from this variant (for example, if the phonological rule says "allow deletion of final [d]", the marginal impact would be measured on the set of utterances in the test database for which the canonical pronunciation ends in [d])

The first metric is significant as it responds to a global, average behaviour, which we don't want to deteriorate, although we could tolerate a certain minimum degradation in system performance, if this overall degradation would be beneficial in some other way. The second metric could seem irrelevant due to its limited impact in the overall performance, but it is justified as it allows the system to correctly recognize a full set of speakers or pronunciations that, otherwise, would bee poorly handled. This population can be small in size, but we must deal with them as in publicly deployed speech recognition systems, it is of outmost importance to reduce the number of speakers for which the system would not work at all. In order to calculate this marginal improvement, we need to know in advance the subset of the whole database than can be improved.

So, our proposal consists of the following stages:

1. Apply the selected phonological rules to the canonical task dictionary, generating the "modified" task dictionary
2. For every phonological rule:
   - Calculate the subset of the test database that can be benefited by the application of the rule
   - Run a recognition experiment on the calculated subset, evaluating (with specific metrics) the differences between using the canonical dictionary and the modified one (with added variants)

The recognition experiments could be done using the overall recognition system described in section 2.1, the hypothesis module, or the verification module. The general idea is based on checking the output of any speech recognition module and calculating in which position within the (sorted) list of recognized words, the correct word was actually recognized (let's call this value "recognized position": RP_c when using the canonical dictionary and RP_m when using the modified one). After making this calculation using the canonical and modified dictionaries, we can compare the results and decide, for every utterance, which dictionary got the best result (the one with the lower RP). The higher the differences between RP_c and RP_m, the higher the impact of the phonological rule applied.

3.2. Marginal improvement metrics

In [7] we include a full list of the evaluation metrics we proposed (19 different ones) but, for the sake of brevity, we will show just a subset of them here:

- Number of utterances for which there was no difference between both dictionaries ($RP_c = RP_m$)
- For the utterances in which the canonical pronunciation was preferred ($RP_c < RP_m$):
  - Number of utterances
  - Average difference between RP_c and RP_m
  - Relative difference between RP_c and RP_m (measured as a percentage relative to RP_m)
- For the utterances in which the modified pronunciation was preferred ($RP_c > RP_m$):
  - Number of utterances
  - Average difference between RP_m and RP_c
  - Relative difference between RP_m and RP_c (measured as a percentage relative to RP_c)

Each of the proposed metrics gives more insight on specific aspects of the impact of the added variants (overall impact, absolute or relative gain, details on canonical or modified dictionaries improvements and preferences, etc.)

4. Experimental results

4.1. Search space increase

The relative increase in dictionary size (compared to the canonical one) when applying the *dfinal* rule is between 10% and 15% for the two task dictionaries, while this increase rises up to over 30% if we apply *all of 12 selected rules* mentioned in section 2.3.

4.2. Overall impact in error rate

When evaluating the overall impact in performance using the *hypothesis* module and the 12 phonological rules, we found a not statistically significant relative increase in error rate under 0.89%.

The same comments can be made when we use the *verification module* (which is much more powerful in terms of recognition accuracy: the baseline error rate of the hypothesis module is 51% and that of the verification module is below 15%), with a not statistically significant relative error rate increase under 4.4%.

The results show that the overall impact in the global error rate is far from being statistically significant, especially if we consider that the increase in the dictionary size was as big as 30%. Nevertheless, these results are consistent with others shown in the literature and, additionally, we must consider that it's very difficult to notice the global effect of the use of phonological variations, mainly because, in most cases, there are very few examples that can be benefited from them.

4.3. Marginal impact evaluation

In table 4 we show, as an example, the results for the application of the dfinal rule (word ending [d] deletion) in the hypothesis module with the PERFDV database.

Table 4*: Marginal improvement results* when applying the *dfinal* rule (hypothesis module)

Metric	Value
Number of utterances for which canonical = modified ($RP_c = RP_m$)	2,313 (92.45%)
Number of utterances for which canonical is better ($RP_c < RP_m$)	173 (6.91%)
Number of utterances for which modified is better ($RP_c > RP_m$	16 (0.64%)
Average (RP_c -RP_m) when canonical is better (and relative value)	1.38 (7.03%)
Average (RP_m -RP_c) when modified is better (and relative value)	58.75 (60.15%)

The first important observation is that the number of words adversely affected by the added variants is much bigger than the number of benefited ones (173 vs. 16). From these figures, it's clear that the application of that rule will never have a statistically significant impact in overall performance and, if any, it will very probably be negative.

If we then focus on the average differences between RP's, we see that the improvement obtained with the modified dictionary is significantly larger than the improvements obtained with the canonical dictionary (58.75 positions vs 1.38, respectively). The relative averages also show this effect: every adversely affected word, in the average, loses 7% in relative position, while the benefited ones gain up to 60%.

These figures clearly show that the marginal negative impact of the rule application is much smaller than its marginal positive impact in the words whose recognition is actually improved.

The same study has been performed for all the selected rules (individually and considering the whole set), with similar quantitative results. As a final example, in Table 4 we include a couple of metrics when using the list of selected rules. The differences in this case are big, but not as important as the ones showed in Table 4, but we have to take into account that the increase in dictionary size is much bigger.

Table 5*: Marginal improvement* results when applying the *selection of rules* (hypothesis module)

Metric	Value
Relative average (RP_c -RP_m) when canonical is better	22.61%
Relative average (RP_m -RP_c) when modified is better	68.31%

This detailed marginal analysis experimentation has also been done using the verification module (on the LOO databases), with similar results (large gains when the modified dictionary performs better), as shown in Table 6.

Table 6*: Marginal improvement* results when applying the *selection of rules* (verification module)

Metric	Value
Number of utterances for which canonical = modified ($RP_c = RP_m$)	9,099 (93.27%)
Number of utterances for which canonical is better ($RP_c < RP_m$)	548 (5.62%)
Number of utterances for which modified is better ($RP_c > RP_m$	109 (1.12%)
Average (RP_c -RP_m) when canonical is better (and relative value)	3.4 (41.06%)
Average (RP_m -RP_c) when modified is better (and relative value)	12.7 (78.56%)

If we use the verification module, it's worth mentioning that we are mainly interested in checking

what happens with the words recognized in the top of the recognized list, as this will be the final system output. In that case, when we introduced the variants, the number of utterances that loses the first position is 114 and the ones gaining this position is 53, which means a net loss of 76 files, less than 0.4% of the available database. The important fact is that those 53 "improved" files would have had no possibility of being recognized without the use of pronunciation variants, while the net loss is not statistically significant.

5. Conclusions and future work

In this paper we have presented an alternative proposal on the evaluation of pronunciation variations, that consists of taking also into account the marginal improvements achieved by the introduction of such variants.

We have defined specific marginal improvement metrics, specially adapted to get more insight on the actual impact of the pronunciation alternatives used, from different perspectives, and performed an experimental evaluation on a real-world isolated word telephone speech task using a versatile speech recognition system based on the hypothesis-verification paradigm.

Our results show that, even though the overall impact of the pronunciation alternatives is not statistically significant (even in dictionaries significantly bigger than the canonical ones), we can get marginal benefits, important enough for certain users or pronunciations to have an opportunity to be correctly recognized. While it's true that in our strategy we are explictly ignoring part of the added confusion due to the inclusion of new variants (restricting the evaluation to the words 'affected' by the rule application), our approach should be interpreted as one step to gain further information on the mechanisms leading to successful rule generation and evaluation. Future work will be devoted to researching how these marginal improvements on the subsets can be exploited without loss on the rest of the data.

We are also working in evaluating this approach on continuous speech recognition tasks, and using the marginal improvements metrics as figures-of-merit to assess the introduction of pronunciation variants in speech recognizers, which is one of the more active areas in pronunciation variation modeling these days. Another open research line is developing a methodology to apply our proposal to multiple pronunciation strategies using a data-derived approach.

6. Acknowledgements

This work has been done under the contract TIC2003-09192-C11-07 MIDAS-INAUDITO with the Comisión Interministerial de Ciencia y Tecnología to whom we want to thank for their support. We also want to thank the rest of the Speech Technology Group for their collaboration and support.

7. References

[1] Strik, H. and Cucchiarini, C. "Modeling pronunciation variation for ASR: a survey of the literature". *Speech Communication*, vol 29, p. 225-246. 1999.

[2] Saraçlar, M., Nock, H. and Khudanpur, S. "Pronunciation modeling by sharing Gaussian densities across phonetic models". *Computer Speech and Language, vol 14, p. 137-160*. 2001.

[3] Wester, M. "Pronunciation modeling for ASR – knowledge-based and data-derived methods". *Computer Speech and* Language*, vol 17, p 69-85*. 2003

[4] Holter. T. "Maximum Likelihood Modelling of Pronunciation in Automatic Speech Recognition". *PhD. Thesis. Norwegian University for S&T*. 1998.

[5] Riley, M., Byrne, W et al. "Stochastic pronunciatio modeling from hand-labelled phonetic corpora". *Speech Communication, vol 29, p. 209-224*. 1999.

[6] Kessens, J. and Wester, M. "Imrpving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation". *Speech Communication, vol 29, p.193-207*. 1999.

[7] Macías-Guarasa, J. "Arquitecturas y métodos en sistemas de reconocimiento automático de habla de gran vocabulario" (Architectures and methods in large vocabulary automatic speech recognition systems). *PhD. Thesis. Universidad Politécnica de Madrid*. 2001.

[8] Tapias, D., Acero, A., Esteve, J., and Torrecilla, J.C. "The VESTEL Telephone Speech Database". ICSLP 94, pp. 1811-1814. 1994.

[9] Strik, H. "Pronunciation adaptation at the lexical level". *Proc. of the ISCA Tutorial & Research Workshop (ITRW) 'Adaptation Methods For Speech Recognition', Sophia-Antipolis, France, pp. 123-131*. 2001

[10] Ferreiros,J. and Pardo,J.M. "Improving continuous speech recognition in Spanish by phone-class SCHMMs with pausing and multiple pronunciations. *Speech Communication 29, pp 65-76*. 1999

[11] Ferreiros, J., Macías-Guarasa, J and Pardo, J.M. "Introducing Multiple Pronunciations in Spanish Speech Recognition Systems". *ESCA Tutorial and Research Workshop on "Modeling Pronunciation Variation for Automatic Speech Recognition", Kerkrade*. May 1998.