INTRODUCING MULTIPLE PRONUNCIATIONS IN SPANISH SPEECH RECOGNITION SYSTEMS

Javier Ferreiros, Javier Macías-Guarasa, José M. Pardo, Luis Villarrubia (*)

Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica.
E.T.S.I. Telecomunicación - Universidad Politécnica de Madrid
Ciudad Universitaria, s/n. 28040 - Madrid. SPAIN.
jfl@die.upm.es
(*) Telefónica Investigación y Desarrollo

ABSTRACT

Pronunciation variations are common sources of recognition errors in real-world applications, so that specific techniques must be developed to handle them. We are describing a method to incorporate pronunciation alternatives that have been tested with both continuous and isolated word speech recognisers for Spanish.

We present an automatic grapheme-to-phoneme system, modified to generate alternate pronunciations. It works according to phonological rules manually developed using certain variations, well known in the linguistic community but not widely exploited in the Spanish speech recognition arena.

We will apply this strategy only to the recognition stage of both a continuous speech recogniser for clean speech data, and an isolated one for a telephone environment task. We will report improvements up to 20% decrease in error rate, for the continuous speech task, while for the isolated word recognition task, no significant effect has been found. We will conclude analysing which effects have led to these results and discuss future work to be done.

1. INTRODUCTION

One of the major difficulties in speech recognition systems is the variability in the speech data, due, among other reasons to alternate pronunciations of words, even within the same speaker [7].

The lexicon is usually composed of a set of words and a single pronunciation for each of them. This pronunciation is considered to be the "standard" or "correct" one, but it usually doesn't have to do very much with the actual pronunciation of the word in a real environment [8].

A first approach is the generation of pronunciations by hand, or even better, applying phonological rules to introduce pronunciation variability. This has been reported to outperform systems using canonical pronunciation forms [8][9]. We will apply this method for both isolated word and continuous speech recognition tasks, showing very different effects on performance.

2. INTRODUCING MULTIPLE PRONUNCIATIONS

To obtain the bootstrap transcriptions of the words in the lexicon into acoustic units, we first developed a rule-guided tool to automati-

cally generate standard pronunciations from the orthographic forms. It can also deal with some coarticulation and assimilation effects in word boundaries, to be used to transcribe training material for continuous speech recognition.

Traditionally, this is considered to be accurate enough for languages such as Spanish, in which standard pronunciations are easy to generate, and the acoustic models can cope, to a certain degree, with pronunciation variability. For languages such as English, this is not generally the case, so that extensive use of hand generated dictionaries has to be used.

Using this standard transcription, we have applied a set of phonological rules to introduce pronunciation variabilities for each word in the lexicon. In the first version of our system, these rules are manually developed, according to expert linguistic knowledge. They address variations commonly found in the castilian Spanish speaking community. To limit the increase in lexicon size, we have reduced the rules to a minimum number, considering those that led to significant improvements in continuous speech recognition accuracy. For isolated word recognition, we started trying the same rules, and finaly excluded some of them according to their actual effect in performance.

3. EXPERIMENTAL SETUP

We are applying the strategy explained above to two different tasks:

◊ **Continuous speech recognition**, for speaker dependent data:

- **Domain**: Access to a navy information database
- **Training set size**: 600 phrases per speaker. **4 speakers**: 2 male and 2 female
- **Test set size**: 100 phrases per speaker, uttered by the same 4 speakers
- **Base dictionary size**: 979 words (standard transcriptions only).
- **Extended dictionary size**: 1211 words (adding multiple pronuntiations, resulting in a 23.7% increase in dictionary size).

◊ Isolated word recognition, for speaker independent data in telephone environment.

- **Domain**: Proper names (first and second names, surnames): Vestel database [1]
- **Training set size**: 5800 words
- **Test set size**: 2500 words
- **Base dictionary size**: 1175 words
- **Extended dictionary size**: 1266 words (applying the same rules than in the CSR case, resulting in a 7.7% increase in dictionary size) and 1193 words (excluding the rules related to transformations of initial *b, d* and *g*, resulting in just 1.5% increase in dictionary size).

4. CSR SYSTEM ARCHITECTURE

The CSRS is a one-pass algorithm that obtains the best word sequence without the help of any grammar. Words are transcribed into allophone sequences and in this point is where pronunciation variations are introduced for some words [5].

Acoustic Processing (AP): The input speech signal is preprocessed obtaining 11 basic (10 MFCCs + Energy), 11 delta and 11 delta^2 parameters. They are quantized for discrete HMMs (DHMMs) or soft quantized if semi-continuous HMMs (SCHMMs) are used, with 3 codebooks and 256 centroids each.

Continuous speech recognition system (CSRS): it is a one-pass algorithm that obtains the best word sequence without the help of any grammar. Words are transcribed into allophone sequences and in this point is where pronunciation variations are introduced for some words [5]. We have an inventory of 47 basic units plus 2 silence units and a unit to improve the modelling of the acoustic events between words. The 47 basic units include special models for stressed, nasalised and stressed nasalised vowels, two HMM models to form the stops (the closure and the burst), two models for the Spanish /ñ/ sound.

5. IWR SYSTEM ARCHITECTURE

The IWRS is actually a pre-selection system [2][4][6] to be used as a hypothesis subsystem, to be run before a detailed match recogniser. It follows a bottom-up, two state strategy, and its main modules are described below: Acoustic Processing (AP), Phonetic-string build-up (PSBU) and Lexical Access (LA) [3].

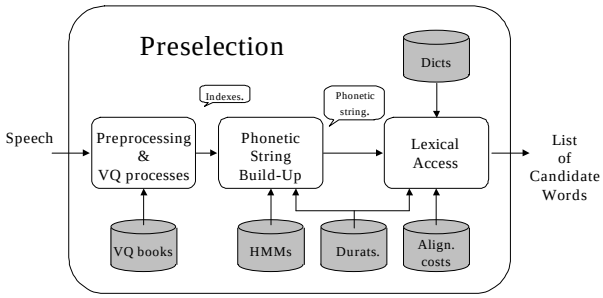


Figure 1. IWR Architecture

Acoustic Processing (AP): The input speech signal is preprocessed (8 MFCCs, 8 delta-MFCCs, cepstral energy and its first derivative) and quantized for discrete HMMs (DHMMs) or soft quantized if semi-continuous HMMs (SCHMMs) are used, with up to 2 codebooks and 256 centroids each).

Phonetic String Build-Up (PSBU): the resulting indexes are passed to the phonetic string build up module which generates a string of alphabet units. We have used the One-Pass algorithm with minor modifications. We have used 25 allophone-like units that have been automatically extracted using an HMM clustering algorithm. The initial allophone inventory was composed of 51 units and 2 of the models are modelling initial and final silences. We have kept the number of units so low to reduce the computational complexity required for the task. Nevertheless, we have also tested 51 units inventory, without getting significant improvements.

Lexical Access (LA): The phonetic string is matched against the dictionary, using a dynamic programming algorithm and alignment costs for unit substitution, insertion and deletion errors [11]

Additional tools were designed for the training stages, dictionary handling, grapheme-to-phoneme conversion, automatic HMM clustering, database analysis, etc.

6. RESULTS

Continuous Speech

In the next page, you can see Table 1 and Figure 2, where we show the results obtained for discrete context independent HMMS (**dd**), discrete contextual HMMS (**ddcn**), semicontinuous context independent HMMS (**sc**) and semicontinuous phoneme dependent contextual HMMS (**scpd**). In every case we include word accuracy rates using normal dictionaries, dictionaries with multiple pronunciations and the corresponding error rate reductions.

	Normal	Multiple	Error rate reductions
dd	66.6%	70.8%	-12.6%
ddcn	69.3%	72.9%	-11.8%
sc	75.3%	78.7%	-13.8%
scpd	80.0%	83.9%	-19.6%

Table 1. Results for the CSR tests

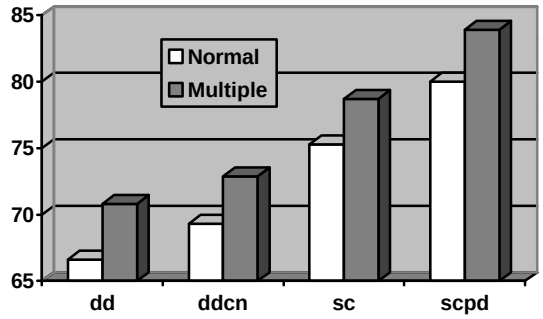


Figure 2. Results for the CSR tests

In Table 2 and in Figure 3, we include the results for the discrete (**dd**) and semicontinous (**sc**) context independent HMMs. The first column refers to the normal dictionary (1175 words), the second to a dictionary with all the rules for multiple pronunciations (1266 words) and the third excluding the rules related to *b d* and *g* initial variants (1193 words). In the later case, dictionary size increased only 1.5%. We decided to test these two cases, as the transformations related to initial *b d g*, were not supposed to report any benefit in recognition rate for the isolated case, compared to continuous speech recognition, where these rules have actually shown to be really useful, in order to correct limitations in the grapheme to phoneme converter. Anyway, and as can be clearly seen, no significant improvement was achieved, so that we don't include error rate variation in this case.

	1175w	1266w	1193w
dd	72.6%	72.5%	72.6%
sc	82.5%	82.2%	82.2%

Table 2. Results for the IWR tests

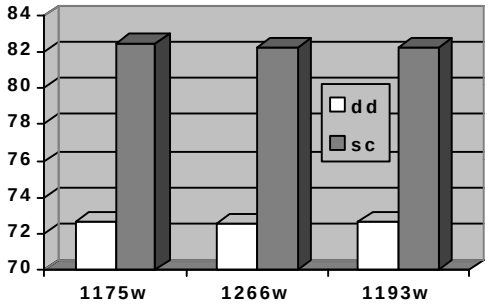


Figure 3. IWR Architecture

7. CONCLUSIONS

Through the results obtained, we can clearly see that the use of multiple pronunciations had a direct and significant impact on the continuous speech recognition system.

We think that in this case, the fact that it allows the dynamic programming algorithm to chose the right beginning allophone for certain words, depending on the left context, is the main reason for the improvement achieved. This is very important in castilian Spanish and is mainly solved by one of our rules that proposes both fricative and occlusive variants of initial *b*, *d* and *g* sounds. Roughly speaking, in Spanish, *b*, *d* or *g* preceded by a pause or a nasal sound should be occlusive while they are considered to be fricative in the rest of the cases.

On the other hand, this important rule for the CSR system performance, seems to have no effect in the isolated word tests, because all the words are uttered after a pause, so that the occlusive version applies always. In addition

to this, even the perplexity increase may be one of the causes leading to slightly lower recognition accuracy.

The initial *b d g* variant rule produces the biggest increase in perplexity, both in the isolated and continuous speech tasks. Nevertheless, being a rule that was specifically designed to cope with co-articulation effects in the beginning of the words, little gain can be obtained in the isolated recognition system. If we consider the experiment where we excluded the initial *b d g* variant rule, the little increase in word pronunciation alternatives does not seem to be sufficiently relevant to produce significant effects.

Moreover, in the isolated recognition task, the database includes so many speakers that the dialectal component is supposed to be relevant, while the rules used apply only to castilian Spanish.

8. FUTURE WORK

We still need a careful review of the results in order to know which rules led to pronunciation alternatives that are actually chosen during the recognition process. A bigger database would help to check the behaviour with higher vocabulary sizes as it is useless to simply add new words, perhaps with several variants, but without similar acoustic examples in the test set that could be matched against these variants.

One of the main drawbacks of the initial strategy approach taken is its lack of correlation with actual training data. We will introduce this alternate pronunciation mechanism in the training stage of our recognisers, to improve acoustic model estimation, and to realise till what extent the rules applied are actually present in real data.

We have not taken yet into account well known dialectal variations, as it would involve increasing significantly the dictionary size, and the information would be too specific to be applied in a general real-world testing environment. However, we are planning to add rules to model such variations, and using them in specific database subsets, given the dialectal source of the speech material is known. We will also address the automatic generation of pronunciation baseforms, as the most promising strategy in this field [8][9].

9. REFERENCES

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

[2] M.A. Leandro and J.M. Pardo. *"Low Cost Speaker Dependent Isolated Word Speech Preselection System Using Static Phoneme Pattern Recognition"*. Eurospeech 93, vol. 1, pp. 117-120. 1993.

[3] J. Macías-Guarasa et al. *"On the Development of a Dictation Machine for Spanish: DIVO"*. ICSLP 94, S22-26. pp. 1343-1346. 1994.

[4] J. Macías-Guarasa et al. *"Comparison of Three Approaches to Phonetic String Generation for Large Vocabulary Speech Recognition"*. ICSLP 94, S36-22. pp. 2211-2214. 1994.

[5] J. Ferreiros. *"Contribution to Continuous Speech Recognition HMM Training Methods"*, PhD. Thesis, Nov. 1996.

[6] J. Macías-Guarasa, A. Gallardo, J. Ferreiros, J.M. Pardo and L. Villarrubia. *"Initial Evaluation of a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone*

Environment”. ICSLP 96, pp. 1343-1346. 1996.

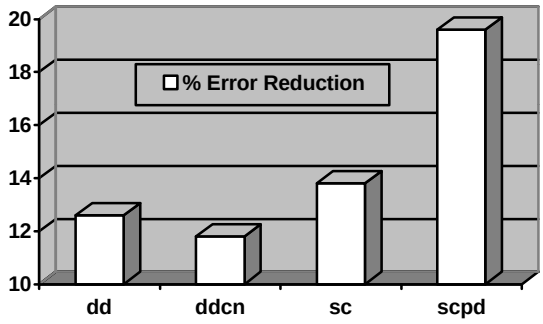
[7] C, Wooters and A. Stolke. *"Multiple-Pronunciation Lexical Modeling in a Speaker Indepent Speech Understanding System"*. ICSLP94, pp. 1363-1366. 1996

[8] C.M. Westendorf, J. Jelitto. *"Learning Pronunciation Dictionary from Speech Data"*. ICSLP 96, pp. 1045-1048. 1996

[9] P. Schmid, R. Cole, M. Fanty. *"Automatically Generated Word Pronunciations from Phoneme Classifier Output"*. ICASSP 93, vol 2. pp 223-226. 1993

[10] Xabiert Aubert and Christian Dugast. *“Improved acoustic-phonetic modelling in PHILIPS´ dictation system by handling liasions and multiple pronunciations”*. Eurospeech 95, pp 767-770.

[11] Fissore, L., Laface, P., Micca, G. and Pieraccini, R. *“Lexical Access to Large Vocabularies for Speech Recognition”*. IEEE Trans. ASSP Vol. 17, n. 8. 1197-1213. 1989


20
18
16
14
12
10
% Error Reduction
dd
ddcn
sc
scpd