Pronunciation modeling using finite state transducers

I. Trancoso[*], **D. Caseiro**[*], **C. Viana**[†], **F. Silva**[*] and **I. Mascarenhas**[†]

[*] INESC-ID/IST
Rua Alves Redol 9, 1000-029 Lisbon, Portugal

[†] CLUL
Av. Prof. Gama Pinto 2, 1749-016 Lisbon - Portugal

ABSTRACT

This paper describes some experiments with pronunciation modeling for spontaneous speech in European Portuguese. The transducer framework provides an elegant way to combine a pronunciation lexicon of canonical forms with alternative pronunciation rules. The main phonological aspects that the rules are intended to cover are: vowel devoicing, deletion and coalescence, voicing assimilation, and simplification of consonantal clusters, both within words and across word boundaries. Our aligner proved sufficiently robust to be able to process fairly long dialogs with overlapping turns, despite many limitations, namely in terms of absence of models for voice quality changes.

1 INTRODUCTION

This paper describes our efforts at modeling pronunciation variation via rules expressed through finite state transducers. Following the same type of reasoning described in [1], we attempt to measure the effectiveness of our rules through forced alignment, in the context of spontaneous dialogs. Alignment can be of interest for both synthesis and recognition purposes. For instance, we can use it for automatically placing phone labels and boundaries to be used in concatenative speech synthesis using variable length units. And we can use it for aligning large quantities of speech signals and corresponding text in a bootstrap procedure to better retrain our acoustic models in large vocabulary continuous speech recognition.

At the Spoken Language Systems Lab of INESC, we have extensively used this bootstrap procedure, in the development of our broadcast news recognition system (AUDIMUS+) [2]. It is a hybrid speech recognition system that combines the temporal modeling capabilities of Hidden Markov Models (*HMMs*) with the pattern discriminative classification capabilities of multilayer perceptrons (*MLPs*). The system combines context-independent posterior phone probabilities generated by several *MLPs* trained on distinct feature sets resulting from different feature extraction processes. All *MLPs* use the same phone set constituted by 38 phones for Portuguese plus silence and breath noises.

The initial acoustic models have been trained with a very small corpus of manually segmented data and retrained through several alignment steps using large quantities of BN data. As a result, the acoustic models which we are also using in this work exhibit a considerable generalization capability. After several alignment steps, Audimus+ has now reached an error rate of 14.8% in F0 conditions at 7.6 xRT, which contrasts with 28.8% for spontaneous speech. But just adding more transcribed data and more alignment steps will not contribute any longer to significant improvements.

This explains the motivation for following this line of work. Rather then doing alignment experiments with spontaneous speech segments of broadcast news, we decided to do them with the only corpus of spontaneous speech for which we have manual segmentation at the phone level - the Coral dialog corpus.

All the alignment experiments described here have been done in the framework of weighted finite state transducers (*WFSTs*). In fact, we use the *WFST* framework in AUDIMUS+, taking advantage of the elegant and uniform formalism that allows very flexible ways of integrating multiple knowledge sources, and the superior search performance obtained when the search network is optimized using automata determinization and minimization. The adaptation of our decoder to do forced alignment was first motivated by the need to align digital spoken books at a word level [3]. The success of this experiment - the alignment of a 2h15m audio file ran at 0.03 xRT, excluding acoustic modeling - motivated us to try a similar procedure for fairly long dialogs, containing many overlapping turns.

This paper thus has 5 main parts, described in the following sections: section 2 describes the Coral corpus, its annotation and some relevant statistics; section 3 discusses the way we dealt with pronunciation variation; section 4 describes our aligner and the syntax of alternative pronunciation rules; experimental results are shown in section 5; finally, section 6 summarizes the main conclusions of this work.

ISBN 1-876346-nn-n

2 THE CORAL CORPUS

Coral is a map task dialog corpus, involving spontaneous conversations between pairs of speakers about map directions. In the 16 different pairs of maps, the names of the landmarks were chosen to allow the study of some connected speech phenomena: sequences with /l/ favoring or not its velarization (e.g. *sala malva, sal amargo*); sequences with /s/ in word final position followed by another coronal fricative (e.g. *poços secos*); sequences of plosives formed across word boundaries (e.g. *clube de tiro*); and sequences of obstruents formed within and across word boundaries (e.g. *bairros degradados*).

The recordings involved 32 speakers, and took place in a small sound proof room. The two speakers were separated by a small screen wall, whose goal was to avoid direct visual contact, but did not provide acoustic isolation. The speakers wore close-talking microphones and the recordings were made in stereo directly to DAT and later down-sampled to 16 kHz per channel.

Given the recording conditions, a reasonable amount of cross-talk from the other channel was clearly audible. Our first alignment experiments were made with the original signals and produced too bad results - the end of the turns was not properly detected, which caused words from one speaker to be frequently aligned during the other speaker's turn. The problem was aggravated when overlap occurred, which was fairly frequent.

In order to reduce this cross-talk, we adopted an adaptive noise canceling scheme [4], in a symmetric architecture, in order to estimate both source signals simultaneously. Each filter had 256 taps (16ms) and was adapted using the standard LMS algorithm. Only one of the filters was adapted at each time step and no attempt was made to avoid adaptation during overlap periods, since these are usually short and not enough to jeopardize the filter estimates. Using this scheme, an average cross-talk reduction of 10dB was achieved for the weaker interference signal, and of 18dB for the stronger interference [5]. At the perceptual level, the interfering signal became almost inaudible and no loss of quality was observed in the reference primary signal. Given these good results, all the experiments reported in this paper refer only to the signals after channel separation.

All dialogs were orthographically transcribed following the same transliteration conventions using SGML format of other map task corpora[1]. Only a small pilot dialog was annotated at all levels (including, fairly recently, the phone level, using the SAMPA phonetic alphabet, but only for the left channel). Audio samples and corresponding transcription of some turns of this pilot dialog can be found in the group's website.[2]

[1] http://www.hcrc.ed.ac.uk/dialogue/maptask.html
[2] http://www.l2f.inesc-id.pt/projects/coral/ortograf.html

2.1 CORPUS ANALYSIS

Of particular importance to this work is the analysis of the orthographic annotation tags of this corpus which explicitly indicated the adopted pronunciation. Altogether, there were 4076 such tags. 31% mark glide deletion (e.g. [b"aSu] instead of [b"ajSu], for the word *baixo*, meaning down). Another significant percentage (30%) occurs with the prepositions *para* and *por* and their contractions with other function words (e.g. [pO] instead of [p6r6 u]); an interesting case is the word *por*, which was only pronounced in its canonical form [pur] 61% of the times, an alternative form [pru] being also very frequent (32%). Truncation of the initial syllable is also fairly frequent (21% in forms of the verb *estar* (to be), and 5% in other words). 5% simply mark lengthening in monosyllabic function words.

The analysis of the manual phone alignment of the pilot dialog revealed other phenomena not explicitly marked by the annotators in the orthographic transcription. In fact, by comparing the observed pronunciation with the canonical one, we notice deviations in 46% of the forms, 59% of which correspond to function words and auxiliary verb forms. By analyzing these deviations, we see that schwa deletion accounts for the larger percentage (20%), namely at word edges (16%). The deletion of unstressed [u] also accounts for a very significant percentage (18%). Next comes glide deletion and monophtongization, affecting both oral (e.g. aw $\rightarrow$ O) and nasal (e.g. 6 $\sim$ j $\sim\rightarrow$ e $\sim$) diphthongs (5% and 10%, respectively).

Other deviations at word boundaries are also very frequent, affecting namely words ending in [S], when followed by vowels or voiced consonants (8%), coalescence of two [6]'s in successive words, etc. Of particular importance to this work is the analysis of the deviations that result from voicing assimilation and vowel and consonant deletion and coalescence across word boundaries. The examples provided by the landmark names do not differ from our expectations. The simplification of consonant clusters involving two plosives seems to occur more frequently when they are equal or have the same place of articulation. Likewise for the simplification involving two or more coronal fricatives. These phenomena seem to be more frequent within word then across word boundaries.

3 MODELING PRONUNCIATION VARIATION

The way we dealt with pronunciation variation has some similarities with the one described in [6]. The variations that depend on word-level features of lexical items (such as part-of-speech) and those that are particular to specific lexical entries (such as many acronyms in Portuguese, for instance) are just included in the lexicon. We shall denote by *Lex0* the Coral lexicon that includes only canonical forms and multiple

pronunciations for 21 heterophonic homographs.

The remaining variants that depend on the local immediate segmental context are modeled through rules whose syntax will be described in the following section. Rather than specifying rules which would mainly affect function words and forms of the verb *estar* (to be), we included in the lexicon *Lex1* multiple pronunciations for 40 such words, which were so frequently marked with micro-annotations in our corpus.

Some of the rules concern variations that depend on the stress and syllable position. The lexicon uses different labels for representing segments in particular positions. For instance, label *I* denotes a frequent alternation between [i], [E] and [e] in the beginning of some words starting by "e". When no rules are applied, the default pronunciation is [i].

The main phonological aspects that they are intended to cover are: The main phonological aspects that the rules are intended to cover are: vowel devoicing, deletion and coalescence, voicing assimilation, and simplification of consonantal clusters, both within words and across word boundaries. Some common contractions are also accounted for, with both partial or full syllable truncation and vowel coalescence. Vowel reduction, including quality change, devoicing and deletion, is specially important for European Portuguese, being one of the features that distinguishes it from Brazilian Portuguese and that makes it more difficult to learn for a foreign speaker. As a result of vowel deletion, rather complex consonant clusters can be formed across word boundaries, as mentioned in section 2.

4 ALIGNMENT

Our aligner is based on *WFSTs* in the sense that its search space is defined by a distribution-to-word (or distribution-to-phone) transducer that is built outside the decoder. For the alignment task, that search space is usually build as $H \circ L \circ W$, where H is the phone topology, L is the lexicon and W is the sequence of words that constitutes the orthographic transcription of the utterance. As no restrictions are placed on the construction of the search space, it can easily integrate other sources of knowledge, and can be optimized and replaced by an optimal equivalent one.

In order to cope with possible de-synchronizations between the input and output labels of the *WFST*, the decoder was extended to deal with special input labels that are internally treated as epsilon labels (similar to skip arcs in *HMMs*), but are used to mark time transitions or boundaries. Whenever such end-of-segment labels are crossed, the time is recorded in the current hypothesis. The user may choose to place those labels at the end of each phone *WFST* or at the end of each word *WFST*, depending on choosing either phone-level or word-level alignment, respectively.

4.1 ALIGNMENT WITH ALTERNATIVE PRONUNCIATION RULES

When aligning using alternative pronunciation rules, the search space becomes $H \circ R^{-1} \circ L \circ W$, where R^{-1} is the inverse of the rule transducer.

The rules are specified using a finite-state grammar whose syntax is similar to the Backus-Naur-form augmented with regular expressions. We added the operator $\rightarrow$, simple transduction, to the usual set of operators, such that $(a \rightarrow b)$ means that the terminal symbol a is transformed into the terminal symbol b. The language allows the definition of non-terminal symbols (e.g. *$vowel*). All rules are optional by default, and are compiled into *FSTs*. In our case, we did not have enough manually labeled material to train weights. We do not apply the rules one by one on a cascade of compositions, but, because they are optional, we rather build their union in order to avoid the exaggerated growth of the resulting transducer, which can be exponential with the length of the composition cascade. The rule transducer R is thus build as $R = \bigcup_i \Sigma^*(R_i\Sigma^*)^*$ where R_i is the transducer corresponding to a particular rule specification expression.

We also allow the specification of negative constraints, or *forbidden* rules, that disallow the occurrence of expressions or sequences. Such an expression R is compiled to $\Sigma^* \cap \overline{\Sigma^* R \Sigma^*}$.

Figure 4.1 shows an example of a sandhi rule specification, together with the forbidden counterpart. The rule set allows for /S/ not to be changed into /z/ only when there is a silence before the next word starting by a vowel. If there is no silence, then the path /S/ end-of-word (EOW) vowel is forbidden.

A different type of rules, involving contractions and multi-word reductions can also be implemented through a reduction transducer (Rd) that encodes rules that map such reductions to their canonical form [6] (e.g. *gonna* $\rightarrow$ *going to*). The aligner search space becomes then $H \circ R^{-1} \circ L \circ Rd \circ W$.

$\$V = (\$Vowel | \$NasalVow | \$Glide | \$NasalGli)$
$\$WB = (EOW\ (sil \rightarrow NULL)\ (EOW \rightarrow NULL))$
`DEF_RULE S_z`, $(\$V\ (S \rightarrow z)\ \$WB\ \$V)$
`FORBIDDEN_RULE No_S_z`, $(S\ EOW\ \$V)$

Figure 1: Example of rule specification.

5 EXPERIMENTAL RESULTS

This section describes our alignment results with the small subset of the Coral corpus that has been manually annotated at the phone level. In order to evaluate the distinct versions of our aligner, we used as a metric the phone level error rate and two additional measures: the percentage of matching phone labels for which the absolute error is less than 10ms and the av-

ISBN 1-876346-nn-n

erage absolute error in 90% of the cases. The results are shown in table 1. A dynamic programming algorithm was developed to match the manually annotated labels with the automatically derived ones, minimizing their string-edit distance. The algorithm penalizes substitutions, insertions and deletions (costs 10, 7 and 7 respectively), but favors very common ones (cost 3).

Our first experiment was made with Lex0 and no alternative pronunciation rules. An analysis of the largest errors shows they are due to the fact that we did not try to align laughs, annoying grunts, and filled pauses, which causes severe misalignments in the neighboring words (up to 5 ms). In fact, our acoustic models could not yet cope with such phenomena. The performance in overlapping turns is on the same level as the one in non overlapping turns. The second experiment was made with Lex1 and still no rules. The values obtained with this new lexicon were good enough to make all further tests using this lexicon.

We then followed an exhaustive process of testing the efficiency of several types of alternative pronunciation rules. The results obtained with 32 rules are shown in the last line of the table. We were expecting greater improvements, but we cannot dismiss the generalization capabilities of our acoustic models and also the fact that they cannot adequately model laughs and other voice quality changes that seriously affect some portions of the dialogs.

A last experiment was made to test the efficiency of reduction rules developed to handle a couple of examples (e.g. *para a* $\rightarrow$ *pá*) whose reduced forms are rarely lexicalized in Portuguese. These examples, however, were not so frequent in our test corpus, which justifies that the results were practically the same with and without the reduction transducer.

Lex	Rules	%ACC	$\leq$ 10ms	Percentil 90%
Lex0	no	70.04	46.54	0.0122
Lex1	no	71.50	47.29	0.0118
Lex1	yes	78.19	48.57	0.0115

Table 1: Alignment results.

6 CONCLUDING REMARKS

This paper described our first steps toward the study of spontaneous speech in dialogs. We started by characterizing our corpus and explaining how we used channel separation for dealing with the cross-talk in stereo recordings. Our *FST*-based aligner proved sufficiently robust to be able to process fairly long dialogs with overlapping turns, despite many limitations, namely in terms of the absence of models for voice quality changes that are so frequent in this corpus.

The automatic alignment of this corpus is really a crucial step for the study of spontaneous speech phenomena which will have a great impact on the performance of our Broadcast News recognition system. Retraining with automatically aligned material using our pronunciation rules is thus one of the tasks that we are planning for the near future. However, we still have to test the impact of our phonological rules on the recognizer's complexity.

Much remains to be done in the study of pronunciation variation. Some recent studies investigate the relation between sequential pronunciation variants with syllabic restructuring [7]. The use of higher level prosodic information is is also an area which we plan to explore.

ACKNOWLEDGMENTS

The authors would like to thank our colleagues Hugo Meinedo and João Paulo Neto for their prompt cooperation. INESC-ID Lisboa had support from the POSI program of the "Quadro Comunitario de Apoio III".

REFERENCES

[1] N. Cremelie and J-P. Martens, "In search of better pronunciation models for speech recognition," *Speech Communication*, vol. 29, pp. 115–136, 1999.

[2] H. Meinedo, Caseiro D, J. Neto, and I. Trancoso, "Audimus+ a broadcast news speech recognition system for the european portuguese language," in *Proc. PROPOR '2003*, Algarve, Junho 2003.

[3] D. Caseiro, H. Meinedo, A. Serralheiro, I. Trancoso, and J. Neto, "Using wfsts for aligning spoken books," in *Proc. HLT 2002 - Human Language Technology Conference*, San Diego, California, March 2002.

[4] B. Widrow, J. Glover, J. McCool, C. Williams, R. Hearn, J. Zeidler, Dong E., and R. Goodlin, "Adaptive noice canceling: Principles and applications," *Proceedings of the IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.

[5] D. Caseiro, F. Silva, I. Trancoso, and C. Viana, "Automatic alignment of map task dialogs using wfsts," in *Proc. ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation*, Aspen, CO, USA, Sept. 2002.

[6] T. Hazen, I. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," in *Proc. ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation*, Aspen, CO, USA, Sept. 2002.

[7] M. Adda-Decker, P. Mareuil, G. Adda, and L. Lamel, "Investigating syllabic structure and its variation in speech from french radio interviews," in *Proc. ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation*, Aspen, CO, USA, Sept. 2002.