Using Coarticulation Rules in Automatic Phonetic Transcription

Arlindo Veiga[1], Sara Candeias[1], Luís Sá[1,2], Fernando Perdigão[1,2]

[1] Instituto de Telecomunicações, Pólo de Coimbra
[2] Universidade de Coimbra – DEEC, Pólo II, P-3030-290 Coimbra, Portugal
{aveiga, saracandeias, luis, fp}@co.it.pt

Abstract. Phonetic transcription of continuous speech databases is an important and necessary task for speech synthesis and recognition. Manual transcription requires well-trained experts and is time-consuming. So, it becomes more and more usual to implement automatic or semi-automatic transcription procedures. However, it is well known that pronunciation variations and coarticulation phenomena involve serious problems for automatic transcribers. This paper describes an approach to automatically label a continuous speech database at phonetic level using orthographic transcription (word level) and a set of pre-built acoustic models to deal with pronunciation variations. These pronunciation variations arise from coarticulation phenomena. Common intra-word and trans-word coarticulation phenomena will be described as well.

Keywords: Automatic phonetic transcription, Coarticulation, Pronunciation variation

1 Introduction

Careful manual transcription of a speech database produces best results comparing to automatic transcriptions. However, it requires a lot of resources and is usually a forbidden option to transcribe a large corpus. In [1] the authors refer more than one hundred times real-time to do manual transcriptions. Automatic or semi-automatic methods are often used to perform this task with accuracies that tend to manual transcriptions after several iterations. However, particularly in continuous speech, multi-pronunciations and coarticulation must be considered in the automatic transcriptions.

Most of the work in this area, use forced alignment to deal with multi-pronunciation. The alignment is achieved using a dictionary with all pronunciation variation and performing an utterance alignment to choose the correct variation. Obviously, this is a good solution for isolated words, but very inefficient to deal with coarticulation phenomena commonly present in a real speech. Other methods, like the DTW technique [2], are referred to produce accurate results, but they are rarely used.

Many proposed approaches to perform automatic phone alignment are based on HMM modelling, e.g. [3]. In [4] a pos-processing stage is used, while in [3] the acoustic models are changed either by modifying its topology or by tying states and

retraining the whole system thereafter to reflect continuous speech problems. Another approach, [5], changes the decoder network according to pronunciation rules.

Forced alignment and some other proposed methods require changing the dictionary, the acoustic model topology or the recognition network. In the present paper we propose a method that produces a pronunciation network for each utterance taking into account pre-determined rules of intra- and trans-word coarticulations.

This research draws on well known linguistic rules to generate pronunciation variants. A brief overview of the coarticulation topic is presented in section 2 with some relevant aspects of the coarticulatory model that resulted from the phonetic inspection of the database. Section 3 describes the automatic transcription system and section 4 presents experiments and results. Finally, section 5 presents some conclusions.

2 Coarticulation

The coarticulation term, in its original sense, refers to phone coalescence phenomena. One consequence of this coalescence is that resulting phones vary according to the characteristics of neighbouring segments. In fact, broadly speaking, the term coarticulation is often used to refer to this variation. The working definition of coarticulation has been consensual and its effects have been described in terms of their type (i.e., anticipatory or right-to-left, versus carryover or left-to-right) as well as their temporal sphere of influence (look-ahead model versus time-locked model) [6]. We considered that the hypothesis of underlying invariance and the reality of surface variability are displayed as a dynamic procedure. Each of that surface variability is called an alternative phone or allophone. The process of coarticulation exposed here is consistent with Functionalist theory of the Portuguese language [7], and Generativist premises [8]. It is also important to stress that our methodological option lies between phonologic distinctiveness and perceptual phonetics. Coarticulation phenomena are often characterised by the degree of pronunciation variation. Since we work on the linearity of utterance, our purpose does not require syllable structure neither diphthong scope. For convenience's sake, our explanation of coarticulation phenomena does not work based on examination of how syntactic and prosodic rules interact.

The two context chosen for this study are intra- and trans-word scope. We both consider the effects between neighbouring segments at the speech level and environments in which a phoneme could undergo changes. Only coarticulatory phenomena that are critical to this particular database are handled, such as deletions, additions and other linking rules. In fact, we have examined epenthesis and prothesis, the crasis and some sound quality transformation as clipping and assimilation.

All these variations are incorporated in a pronunciation model, expressed in the EBNF (Extended Backus-Naur Form) format. EBNF is a notation used to express context-free grammars. In this work two rules were used: (x|y) rule means that x or y occurs alternatively and the [x] rule implies that x is optional (could be omitted). The used phonetic alphabet is SAMPA [9]. Tables 1 and 2 report coarticulation effects related to specific environments.

The intra-word coarticulation phenomena is based on sequences like [V-to-V] and [(V-to-(V-to-)C]. The following table presents common intra-word coarticulation.

Table 1. Common intra-word coarticulation rules. <V> corresponds to vowel, <C> corresponds to consonant and <V1> groups the phones [6 6~ a E e e~ o o~ O u u~].

Phone sequence	Alternative pronunciation	Coarticulation effects	Lexicon examples (words)
a6~	a[i]6~	epenthesis of an *iod*	atr**aem**, c**aem**
@<V1>	([@]|i)<V1>	vowel quality transformation	ver**ea**ção, aér**ea**, l**ea**l, cand**ee**iro, l**eo**a, afogu**eo**u, ald**eo**la, ant**eon**tem, trans**eun**te
6i	([6]|e)i	vowel quality transformation	resp**ei**to, terr**ei**ro
6E	6[i]E	epenthesis of an *iod*	m**ae**stro, isr**ae**lita
6Z	(6|e)[i]Z	epenthesis of an *iod* and|or vowel quality transformation	igr**ej**a, des**ej**a, grac**ej**a
6J	(6|e)[i]J	epenthesis of an *iod* and|or vowel quality transformation	l**enh**a, des**enh**a, nort**enh**a
o@	o([@]|i)	vowel quality transformation	aperfeiç**oe**, desenj**oe**
uo	u([o]|O)	vowel quality transformation	q**uo**tidiano
<C>ulu	<C>[u]lu	syncope of the first phone *u*	séc**ulo**, escrúp**ulo**, estím**ulo**

The trans-word coarticulation phenomena is based on sequences like [-V-to-V(-to-C)-] or [-C-to-V(-to-C)-]. Descriptions with regard to the segment-boundary patterns occurring in a trans-word scope are provided in Table 2.

Table 2. Common trans-word coarticulation rules. <V> corresponds to vowel, while <C> corresponds to consonants. <V2> groups the phones [6 6~ e~ i o~ u u~]. <son.C> represents the sonorant consonantal like [b d g v z Z m n J L r l].

Phone sequence	Alternative pronunciation	Coarticulation effects	Lexicon examples (sentences)
S sp S	(S[sp]) S	crasis of the *S*	a**s** **ch**amas, o**s** **ch**icotes
S sp <V>	((S [sp])|Z|z)	clipping or quality transformation of the *S*	todo**s** **o**s, a**s** **in**formaçõe**s** **a**cerca
S sp <son.C>	((S[sp])|Z|S) sp <son.C>	clipping or quality transformation of the *S* with assimilation	crítica**s** **d**irigidas, recibo**s** **v**erdes,
S sp <son.C>	((S[sp])|Z|S) sp <son.C>	clipping or quality transformation of the *S* with assimilation	crítica**s** **d**irigidas, recibo**s** **v**erdes,
sp @	sp ([@]|i)	vowel quality transformation	**e**scola, **e**statuto,
sp e	sp ([e]|i)	vowel quality transformation	**e**mulsão, **e**vidente
sp o	sp (o|O)	vowel quality transformation	**o**cupa, **o**lhar, **ou**vir
6 sp 6	(6 [sp] 6)|a	clipping of the *6* or crasis of the *66* with vowel quality transformation	Russi**a** **a**firmou, er**a** **a**ssim
6 sp au	(6 [sp] au)|au	clipping of the *6* with vowel reduction or assimilation of the *6*	lev**a** **ao**, **a** **au**la
u sp <V2>	[u[sp]] <V2>	clipping or syncope of the *u*	reembols**o** **an**tecipado, ferr**o** **en**ferrujado, larg**o** **on**de, vej**o** **u**rsos

3 Automatic Phonetic Transcription System

Forced alignment is widely used to convert orthographic transcriptions into phone level transcriptions. This method produces a word network on the fly from a multi-pronunciation word dictionary. Our approach differs from forced alignment in that a network is produced for each sentence, by applying coarticulation rules to the standard phonetic transcription. The utterance is then decoded with this network and the result taken as the phonetic transcription. This method turns out to be more flexible because intra-word and especially trans-word pronunciation variations can be incorporated. The following example shows a sentence in which usual coarticulations may occur. The coarticulation between words "mesmo" and "os" and between "os" and "alunos" is also indicated Notice that the possible variations depend on the word context, and cannot be included in the dictionary as a context dependency.

Table 3. Example of coarticulations

orthographic:	"ontem mesmo os alunos"
phonetic:	"o~t6~j~ sp me Zmu sp u S sp 6lunuS"
EBNF:	"o~t6~j~[sp]m(e|6)[i]Zmu[[sp]u]((S[sp])|z|Z)6lunuS"

The key of our approach is the generation of the EBNF syntax using predefined linguistic rules (as indicated in the previous sections) and a dictionary with standard transcription of all words considered in our database. The EBNF syntax is converted to an equivalent network (hypotheses grid) before decoding.

The database used in this work is the Tecnovoz corpus [10]. It has about 232,000 Portuguese utterances (around 184 hours) with 254 prompts of commands, 200 prompts of natural numbers and 208 prompts of generic sentences. In this work, for model training and testing, only the sentence part was used. The consistency of the orthographic transcription was automatically evaluated, using a confidence measure based on scores derived from two different decoders. Therefore, from the original 39,966 utterances, only 22,627 were used: 20,365 for training (90%) and 2,262 for testing (10%). These 22,627 utterances correspond to almost 31.5 hours of speech. However, the speech variability is low because there are only 208 different prompts with 1,456 different words.

A set of initial acoustic models is needed to perform automatic transcription. Our initial acoustic models (for context free phones) were trained with the HTK toolkit [11]. We consider 37 phones for the Portuguese language, silence (sil) and short pause (sp), totalling 39 models of phones. The 37 phones for the Portuguese language and the model for silence have the same left-to-right topology with 3 emitting states. The short pause is a "tee model" with one emitting state. In the training, the number of mixtures is incremented until reaching 96.

In order to compare forced alignment with our approach that incorporates coarticulation rules, two set of transcriptions were generated and used to train two set of models: with and without coarticulations. The forced alignment transcriptions of the training database generate 1,278,698 context free phones without counting silence and short pauses, while the transcriptions using coarticulation rules produce 1,224,528

phones. The reason for this difference is that some coarticulation rules allow the skipping of phones and short pauses.

4 Results

The quality of the transcription system was tested using an indirect method. It consisted in performing recognition tests with acoustic models trained with and without coarticulation rules in the transcriptions. The results were evaluated using the utterances from the test set and the `HVite` decoder from the HTK toolkit [11]. The input to the decoder is a network consisting in a phone-loop with all phone models placed in parallel to allow any phone sequence. Because this is a unigram network, low accuracies are expected. A phone bigram was also considered.

To generate transcriptions without coarticulation rules, forced alignment was used. In the test database, this generates 141,894 phones (excluding sp and sil) while our approach generates 135,742 phones. To evaluate the performance of both approaches, we use percentage accuracy defined as:

$$\text{Accuracy} = (N - D - S - I) / N. \quad (1)$$

where N is the number of phones in test database, D is the number of deletions, S is the number of substitutions errors and I is the number of insertions, occurring during the recognition task. We use this measure because it considers all errors occurred in the decoding process. The phone recognition results with a unigram network of phones for all the combinations of models and test transcriptions are presented in Table 4.

Table 4. Phone recognition accuracy on the test database.

Transcriptions	Models without coarticulation	Models with coarticulation
Reference without coarticulation	47.31 %	47.73 %
Reference with coarticulation	49.00 %	49.62 %

These results show an advantage in using coarticulation effects in transcriptions. In addition, we tested the quality of the models trained with and without coarticulation. Results also show that models trained with coarticulations have better performances in all scenarios. For example, for the case of phone @, we found that it is deleted more than 40 % of times. The forced alignment transcriptions do not deal with this situation and generate 117,642 transcriptions of @ compared to 68,432 transcriptions using coarticulation rules. This justifies the decrease of deletions errors caused by this phone.

The trained models have low performance compared with current state of art. To obtain more accurate models, discriminative training was used by means of a minimum phone error method. Accuracy of 59.24 % is archived on the test database with five re-estimations and a phone bigram network. The results for the training database are 62.52 % with a unigram and 66.13 % with a bigram.

Despite the restrictions in selecting the utterances, a preliminary observation showed some remaining inconsistent orthographic transcriptions. However, the main propose of this work is to confirm that coarticulation rules produce better transcriptions, which was indeed verified.

5 Conclusion

An approach to automatic transcription of continuous speech in phonetic level is presented. This approach deals with pronunciation variations that arise from coarticulation phenomena. This is done by generating a network with coarticulations rules for each sentence.

Results show a relative accuracy increment of 4.9% by using coarticulations rules in the phonetic transcription generation.

Future developments include the identification of inconsistent orthographic transcriptions, hesitations and non-linguistic events using the created acoustic models.

Acknowledgments: The two first authors acknowledge FCTUC (Arlindo Veiga) and FCT (Sara Candeias, SFRH/BPD/36584/2007) for their scholarships.

References

1 Kawai, H., Toda, T.: An evaluation of automatic phone segmentation for concatenative speech synthesis. In: Proc. ICASSP, pp. 677--680 (2004)
2 Kominek, J., Bennett, C., Black, A.: Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis. In: Proc. Eurospeech, pp. 313--316 (2003)
3 Kanokphara, S., Tesprasit, V., Thongprasirt, R.: Pronunciation variation speech recognition without dictionary modification on sparse database. In: Proc. ICASSP, pp. 764-767. (2003)
4 Sethy, A., Shrikanth, N.: Refined speech segmentation for concatenative speech synthesis. In: Proc. ICSLP, pp. 145--148. USA (2002)
5 Charonnat, L., Vidal, G., Boëffard, O.: Automatic phone segmentation of expressive speech. In: Proc. LREC, ELRA (2008)
6 Hardcastle, W., Hewlett, N.: Coarticulation: Theory, Data and Techniques. Cambridge, United Kingdom: Cambridge University Press (1999)
7 Barbosa, J.: Études de phonologie portugaise. Junta de Investigaçẽes do Ultramar. Lisboa (1965)
8 Mateus, M. H., d'Andrade, E.,: The Phonology of Portuguese. Oxford University Press (2000)
9 Gibbon, D., Moore, R., Winski, R.: SAMPA computer readable phonetic alphabet. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B (1997)
10 Lopes, J., Neves, C., Veiga, A., Maciel, A., Lopes, C., Perdigão, F., Sá, L.: Development of a Speech Recognizer with the Tecnovoz Database. In: Proc. PROPOR, pp. 260--263. Portugal (2008).
11 Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.0. In: Cambridge University Press (2000).