Acquiring Reusable Multilingual Phonotactic Resources

Julie Carson-Berndsen & Robert Kelly

Department of Computer Science
University College Dublin
Belfield, Dublin 4
Ireland
{julie.berndsen, robert.kelly}@ucd.ie

Abstract

This paper presents a fully automatic procedure for acquiring reusable phonotactic resources from syllable annotated data. The procedure makes use of a regular inference algorithm and the acquired resources are stored in a specialised XML representation. The technique is then extended to support acquisition from phoneme labelled data while providing a semi-automatic annotation system assisting user annotations of phoneme labelled data with syllable boundaries.

1. Introduction

One of the pre-requisites for robust and scalable speech technology is the provision of linguistic resources of varying granularity. This paper is concerned with the induction of regularities in the phonotactic domain in order to develop phonological resources for use in multilingual speech technology applications. The use of phonotactic constraints in speech technology has not been fully exploited by standard stochastic approaches to speech recognition. Usually the term phonotactics in speech technology refers to n-grams which are used to predict the occurrence of a sound based on the $n-1$ preceding sounds. While these probabilistic sequences do cater for some phonotactic constraints, an important generalisation is being missed by such an approach whenever the sequences are all restricted in terms of their length to whatever n is defined to be. However, if the syllable domain is chosen as the limiting factor for the phonotactics, then a phonotactic model for a language can be constructed which is complete and can provide many more restrictions on legal combinations of sounds in a language.

A computational linguistic model for speech recognition which avoids this restriction to sequences of some defined (relatively small) length n for possible phonotactic well-formedness constraints that can be captured is the Time Map model originally proposed by Carson-Berndsen (Carson-Berndsen, 1998). This model provides the motivation for the work presented here. Since the model has been described in detail elsewhere (Carson-Berndsen, 1998), we will only present briefly its main phonotactic resource and then discuss how such resources can be developed for other languages in order to the facilitate multilingual speech applications.

2. Phonotactic Automata

As described in section 1. the Time Map model is a computational linguistic model of speech recognition. It employs a phonotactic resource referred to as the Phonotactic Automaton which guides the recogniser in the identification of well-formed syllables. A Phonotactic Automaton for a language is a finite state representation encoding the allowable sound combinations for that language at the syllable level. It is important to note that a phonotactic automaton describes *all* the permissible combinations of sounds which exist in a language, not just those found in a lexicon of the language. Phonotactic automata also include forms which linguists classify as accidental gaps and can therefore be used in speech recognition to distinguish between well-formed and ill-formed syllables.

Since phonotactic automata conform to the principles of finite-state machines the structure of any phonotactic automaton consists of a finite set of states and state transitions with some state designated as the initial or start state and some subset of states designated as accepting or final states. The transitions of a phonotactic automaton are labelled with single sound segments and the allowable sound combinations are modelled by the state-transition structure of the automaton. Given a syllable structure, a phonotactic automaton for a language can determine if the syllable is well-formed, i.e. conforms to the phonotactics, for the language by attempting to trace an acceptance path through its finite-state structure using the segments of the putative syllable as input symbols. If an acceptance path can be traced then the syllable is well-formed, otherwise it is ill-formed. Since phonotactic automata are simply finite-state machines, tracing paths through the state-transition structure can be performed efficiently using finite state tools (Kelly and Carson-Berndsen, 2003). Complete phonotactic automata for both English and German have already been constructed manually for use in the Time Map model. This required that a linguist establish, for example through dictionary lookup, all permissable sound combinations and record them in the finite-state structure. An alternative, fully automatic approach to acquiring phonotactic automata from syllable labelled data is presented in section 3.

Figure 1 illustrates a subsection of a phonotactic automaton for English showing only a subset of the possible sound combinations observed in well-formed syllables. Note that this automaton is nondeterministic with a unique start state (labelled 0) and final states denoted by double circles. Also, the arcs are labelled with SAMPA[1] phoneme symbols used to represent the individual sound segments. Again, the well-formedness of an onset-peak combination is determined by the state-transition structure of the au-

[1] http://www.phon.ucl.ac.uk/home/sampa/

tomaton, thus the combinations /s p l aI n/ and /T r aI/ would be considered well-formed while the combinations /s p l aI p/ and /T aI/ would be considered ill-formed by this automaton.

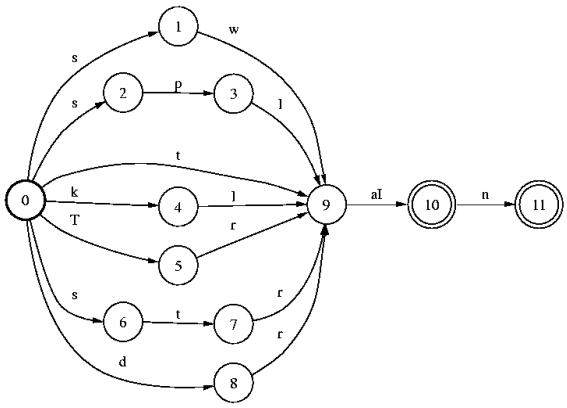

Figure 1: Phonotactic Automaton for English (subsection).

It should be clear that the finite-state model of syllable phonotactics presented in this section satisfies the constraints for a model of phonotactics which accounts for a larger context than that achievable with an n-gram style phonotactics seen, for example, in language identification systems (recently, Adda-Decker et al., 2003). More precisely, phonotactic automata of the type used in the Time Map model describe phonotactics at the syllable level rather than in terms of combinations of two or three segments.

In order to ensure that the phonotactic resources described here can find maximal use in contexts other than the Time Map model, we have designed an XML representation which encodes all the information described by a phonotactic automaton and more. This representation is termed the *Multilingual Time Map* (Aioanei et al., 2004) and the procedure outlined in the following section for automatically acquiring syllable phonotactics has been implemented such that learned phonotactic structures can be exported in this XML representation.

3. Acquiring Phonotactic Automata from Syllable Labelled Data

In order to extend the phonotactic resource catalogue to other languages, we have developed a methodology to support a user in developing phonotactic automata which uses automata induction in line with the approach described in (Carson-Berndsen, 2002). The approach presented here goes significantly further than the work in progress described in that paper in that now a suite of finite state tools with XML interfaces has been implemented which ensures that all resources thus developed can be reused by other interested parties (Kelly and Carson-Berndsen, 2003).

The need for collection and annotation of speech corpora for many languages has long been recognised and recently much emphasis has been placed on multi-level annotations ranging from orthographic word-level annotation to sub-phonetic, feature level annotations. It is this type of multi-level annotation which serves as a basis for the automatic induction of phonotactic resources. This section discusses our *Phonotactic Automaton Learner* (PAL) which takes syllable annotated speech files as input and constructs a minimal deterministic, i.e. canonical, finite state representation of the combinations of sounds contained within the syllables of that data. If only phonemic annotations are available, PAL has been extended to support semi-automatic labelling at the syllable level using incremental learning (see section 4.).

Given a corpus of syllable annotated utterances for some language then the assumption made here is that a corresponding phonotactics can be extracted from the annotations which at least describes the syllable labelled data. Any derived phonotactics is also assumed be an approximation to the complete phonotactics for the language from which the data was drawn. Given such a corpus of syllable annotated data PAL automatically extracts the implicit phonotactics and structures the derived constraints on syllable well-formedness in a finite-state machine. Therefore, PAL must incorporate a regular grammatical inference procedure which infers from a training set of positive examples only, where the regular grammar to be inferred is the phonotactic automaton and the positive training set of examples is the corpus of syllable annotated data. The field of grammatical inference has yielded many important learnability results for different language classes, a full discussion of which is beyond the scope of this paper (a concise summary and discussion can be found in Belz, 2000, Chapter 3). Regarding the inference of phonotactic automata, learnability results state that since the formal language of well-formed syllables in a given natural language is finite, a regular grammar describing the language of syllables (i.e. the required phonotactic automaton) can be inferred from positive data alone.

The choice of regular inference algorithm for PAL is in fact arbitrary, however some inference procedures may be better suited to the specific problem of acquiring phonotactic automata. While future work requires that we experiment with different inference procedures, currently PAL utilises an implementation of the ALERGIA regular inference algorithm (Carrasco and Oncina, 1999). For the purposes of this paper, ALERGIA is illustrated by way of an example. For a more detailed and general discussion of the algorithm as applied to the acquisition of phonotactic automata see (Kelly, 2004). Figure 2 shows a screenshot of the GUI for a multilingual phonological toolkit which has integrated PAL as a central component (Aioanei et al., 2004). A small set of Irish unstressed syllables (Bohan, 2003) is shown in the left hand text area and represents the training sample supplied to PAL. The right hand text area shows a portion of the XML representation of the automaton which has been constructed by PAL using ALERGIA.

ALERGIA first builds a *Prefix Tree Automaton* (PTA) using the sample set of syllables. A PTA is a deterministic finite-state automaton with a strictly branching tree structure that accepts exactly the sample set. A PTA for the first 5 training syllables of figure 2 is shown in figure 3. From the PTA, ALERGIA identifies the canonical automaton through a state merging process whereby two states are merged if the languages described by both states are identi-

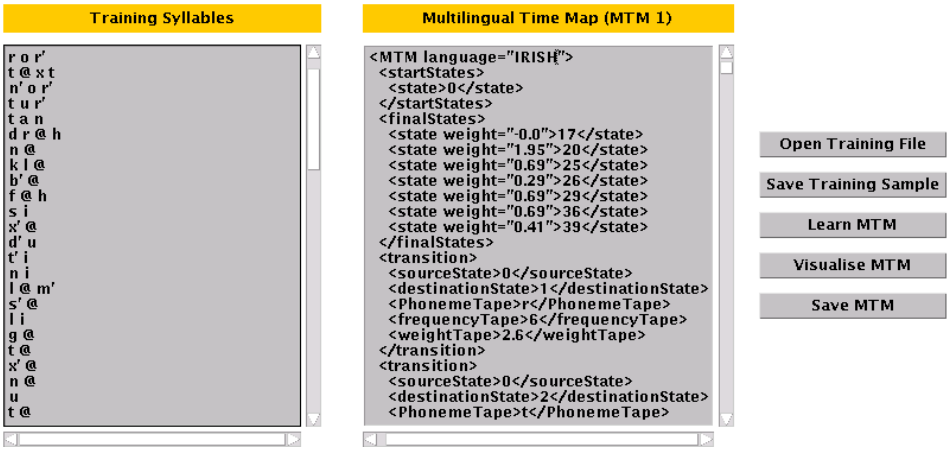

Figure 2: Screenshot of PAL GUI interface.

cal (statistically). Figure 4 shows the canonical automaton derived from the PTA in figure 3.

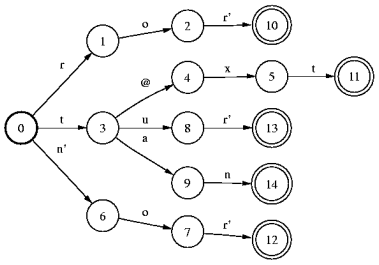

Figure 3: Prefix Tree Automaton.

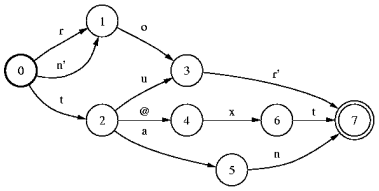

Figure 4: Canonical Automaton.

Since automata are derived from training sets of syllables through the use of a language independent regular inference procedure, PAL describes a generic language independent procedure for acquiring phonotactic resources. However, the procedure is entirely data driven and therefore heavily dependent on the existence and nature of the positive training data. In the first place, a corpus of syllable annotated data may not be available for a particular language especially since corpora are typically labelled at the phoneme and word level but rarely at the syllable level. If a syllable labelled corpus is available then the monetary expense of purchasing the corpus must become a factor. If no such corpus exists then either a new corpus must be compiled and syllable annotated which may be prohibitively expensive and time-consuming. Alternatively syllable annotations can be derived from the annotations provided by some extant corpus (cf. section 4.). A further problem with this approach is that the completeness of the learned phonotactics is dependent on the completeness of the training sample of syllables. Thus, the inferred finite-state structure will only describe information in the syllable domain which is found in the training data. If a rare but valid sound combination does not appear in the corpus then it may never be represented in the acquired phonotactics. However, in order to be complete, the automaton must model all forms which exist in the language in the syllable domain, otherwise it does not represent a significant advantage over n-gram generation (except that it does allow for variable context length). This dependency on training data is usually remedied somewhat by the introduction of generalisation techniques (Neugebauer and Wilson, 2004). Another possible technique whereby the basic onset-peak-coda substructure is imposed on each syllable in the training sample before applying the inference procedure is discussed in (Kelly, 2004). It is anticipated that a combination of these techniques will be required to ensure that inferred phonotactics are adequately complete.

4. Acquiring Phonotactic Automata: Phoneme Labelled Data

This section describes how to obtain phonotactic automata from phoneme labelled data while annotating the utterances of a corpus at the syllable level. This is achieved by extending PAL to PALS (Phonotactic Automaton Learner with Syllabification). PALS assumes that a corpus is annotated at the phoneme level and builds a phonotactics in a semi-automatic incremental manner. PALS iterates through the set of phoneme labelled utterances in a corpus and displays each utterance to a user together with suggested syllable boundaries. Figure 5 for example shows part of an utterance (`Bodan pass ...`) displayed in the PALS GUI with the potential syllable boundaries marked[2].

The suggested boundaries are obtained by using a partial syllable phonotactics that the system has built thus far by observing the marked syllable boundaries in previously

[2]Note that in this case the boundaries have been marked correctly and the result is an additional level of annotation for the utterance in the corpus.

Figure 5: PALS GUI displaying part of a phoneme labelled utterance together with potential syllable boundaries.

processed utterances. Since the phonotactics is partial and incomplete the syllable boundaries may be incorrect and are displayed to the user for verification. This is required in particular for the initial utterances. Thus, for the first utterance encountered, the system will know nothing of syllable structure and the user must mark out all syllable boundaries. A partial phonotactics is then built by applying the ALERGIA algorithm to the syllables marked out for this single utterance. The resulting phonotactics is then used to syllabify the next utterance, however this syllabification will most likely be incorrect and the user must remark the syllable boundaries. Note that this does not necessarily mean that the user must re-syllabify the entire utterance since some of the boundaries for the utterance may be correctly marked. After the second utterance has been syllabified correctly the system rebuilds the partial phonotactics incorporating these new syllables and uses the new phonotactics to syllabify the third utterance. The system continues in this way building up a more complete description of the syllable phonotactics based on the observed syllable boundaries of subsequent utterances. If after a number of utterances the phonotactics is sufficiently complete such that utterances are being marked with correct syllable boundaries then the system can run in automatic mode whereby the syllabification of utterances and the incremental building of phonotactics proceeds without user verification. While this removes the need for user verification, it risks invalid syllable structures in the resulting phonotactics in case the system marks an utterance with erroneous syllable boundaries, which are then incorporated into the phonotactics.

This incremental procedure has been applied to a corpus of Hiberno-English phoneme labelled utterances, known as the HibernoBot corpus.

5. Conclusion & Future Work

This paper has presented an inference procedure for acquiring finite-state phonotactic resources. While the technique is fully automatic in the case of syllable labelled data and supports semi-automatic incremental learning in the case of phoneme labelled data, it is highly dependent on the quality of the corpus annotations. Fortunately, the need for high quality corpora is now recognised and has become an essential part of speech and language technology research. Based on the techniques presented here, investigations are currently under way into the development of tools for validation and consistency checking of corpora.

The phonotactic resources acquired through the inference procedure are specified in XML (as Multilingual Time Maps) to support reusability. There are manifold applications for these acquired resources (Kelly and Carson-Berndsen, 2003) ranging from our own speech recognition application, to language identification or educational applications. The fact that the phonotactic resources are modelled in a uniform way, means that cross-language analyses can easily be undertaken (Carson-Berndsen et al., 2003).

6. Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under Grant No. 02/IN1/I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

7. References

Adda-Decker, Martine, Fabien Antoine, Philippe Boula de Mareuil, Ioana Vasilescu, Lori Lamel, Jacqueline Vaissiere, Edouard Geoffrois, and Jean-Sylvain Lienard, 2003. Phonetic knowledge, phonotactics and perceptual validation for automatic language identification. In *Proceedings of the 15th International Congress of Phonetic Sciences*. ICPhS 03.

Aioanei, Daniel, Julie Carson-Berndsen, Anja Geumann, Robert Kelly, Moritz Neugebauer, and Stephen Wilson, 2004. A multilingual phonological resource toolkit for ubiquitous speech technology. To Appear in Proceedings of the 4th International Conference on Language Resources and Evaluation.

Belz, Anja, 2000. *Computational Learning of Finite-State Models for Natural Language Processing*. Ph.D. thesis, University of Sussex.

Bohan, Amy, 2003. *Synthesising Irish Speech using Temporal Event Representations*. Master's thesis, University College Dublin.

Carrasco, Rafael C. and Jose Oncina, 1999. Learning deterministic regular grammars from stochastic samples in polynomial time. *ITA*, 33(1):1–19.

Carson-Berndsen, Julie, 1998. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. Dordrecht, Holland: Kluwer Academic Publishers.

Carson-Berndsen, Julie, 2002. Multilingual time maps: Portable phonotactic models for speech technology applications. In *Proceedings of the LREC 2002 Workshop on Portability Issues in Human Language Technology*.

Carson-Berndsen, Julie, Ulrike Gut, and Robert Kelly, 2003. Discovering regularities in non-native speech. In *Proceedings of Corpus Linguistics 2003*.

Kelly, Robert, 2004. A language independent approach to acquiring phonotactic resources for speech recognition. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*. CLUK04.

Kelly, Robert and Julie Carson-Berndsen, 2003. Computational linguistic motivations for a finite-state machine hierarchy. In *Proceedings of the 8th International Conference on Implementation and Application of Automata*. Santa Barbara: CIAA 2003.

Neugebauer, Moritz and Stephen Wilson, 2004. Phonological treebanks – issues in generation and application. To Appear in Proceedings of the 4th International Conference on Language Resources and Evaluation.