Phonotactic Constraint Ranking for Speech Recognition

Julie Carson-Berndsen, Gina Joue and Michael Walsh

University College Dublin

Abstract

The aim of this paper is to highlight areas in which a computational linguistic model of phonology can contribute to robustness in speech technology applications. We discuss a computational linguistic model which uses finite state methodology and an event logic to demonstrate how declarative descriptions of phonological constraints can play a role in speech recognition. The model employs statistics derived from a cognitive phonological analysis of speech corpora. These statistics are used in ranking feature-based phonotactic constraints for the purposes of constraint relaxation and output extrapolation in syllable recognition. We present the model using a generic framework which we have developed specifically for constructing and evaluating phonotactic constraint descriptions. We demonstrate how new phonotactic constraint descriptions can be developed for the model and how the ranking of these constraints is used to cater for underspecified representations thus making the model more robust.

1 Introduction

While the success of commercial speech recognition applications has led to a more widespread acceptance of spoken language interfaces, there still seems to be a need for further investigation into the interactions between purely stochastic approaches and more linguistic-symbolic approaches to improve the robustness of multilingual speech systems. The starting point for discussion in this paper is a formally-specified computational linguistic model which has been enhanced by statistical information from various sources to improve the robustness of the model in dealing with the variability of speech and with ’noisy’ input data. Although this paper will concentrate primarily on the extensions to the computational model, we assume also that the fine-grained knowledge representations which are used by the model can be applied to fine-tune stochastic models by providing important underlying structural information (cf. also Jusek et al. (1994)).

The computational linguistic model is the *Time Map* model (Carson-Berndsen 1998) which uses a description of the constraints on the permissible combinations of sounds in a language (phonotactic constraints) to recognise well-formed syllable structures. The phonotactic constraints describe not only those words in the system lexicon but can make predictions as to which words would be considered well-formed by a native speaker of a language. In contrast to stochastic approaches to speech recognition, the *Time Map* model interprets the speech signal in terms of overlap and precedence relations between properties. This allows variability of speech utterances to be modelled by avoiding a segmentation of the speech signal into strictly non-overlapping units. In order to be robust, the model must also cater for imperfect or ’noisy’ input representations and therefore requires a mechanism by which the phonotactic constraints may be relaxed under certain conditions. This paper discusses a methodology for constraint ranking which provides a principled basis for constraint relaxation in the model, based not only on domain knowledge, but also on cognitive factors which influence human production and interpretation. When enhanced by such motivated constraint relaxation procedures, the computational linguistic model will be able to offer insights into how robustness can be addressed in spoken language interface design.

In what follows, we will firstly sketch the *Time Map* model within a generic development environment which facilitates the extension of the technology to other languages (in particular languages which have received little attention thus far) and feature systems. Secondly, we will introduce the notions of constraint relaxation and output extrapolation as assumed by the model and discuss how these mechanisms are employed using a ranking of the constraints. We then discuss how the constraint ranking is achieved based on a functional cognitive analysis of phonological data. The paper concludes with some references to further developments with respect to the extension of the language functionality.

2 LIPS and the Time Map Model

The *Time Map* model was proposed as a computational linguistic model for speech recognition by Carson-Berndsen (1998) and has been tested within a speech recognition architecture for German. The model has recently been extended to English and has been provided with an interface which allows users to define and evaluate phonotactic descriptions for other languages and sublanguages. This generic development environment is known as the Language Independent Phonotactic System (Carson-Berndsen and Walsh 2000a). LIPS aims to provide a diagnostic evaluation of the phonotactic descriptions in the context of speech recognition. That is to say, rather than just providing recognition results, partial analyses can be output indicating which constraints have or have not been satisfied and where the parsing breaks down. This, together with the constraint relaxation and output extrapolation procedures to be discussed below, allows adequate parameters to be chosen that define a compromise between maximal recognition rates and minimal analysis overhead.

The *Time Map* model uses a finite-state network representation of the phonotactic constraints in a language, known as a phonotactic automaton, together with axioms of event logic to interpret multilinear representations of speech utterances. A subsection of a phonotactic automaton for CC- combinations in English syllable onsets can be seen in figure 1 and an example multilinear representation of a simple single word utterance in figure 2. The transitions in the phonotactic automaton define constraints on overlap relations which hold between features in a particular phonotactic context (i.e. the structural position within the syllable domain). [1]

[1] The monadic symbols written on the arcs in figure 1 are purely mnemonic for the feature overlap constraints they represent.

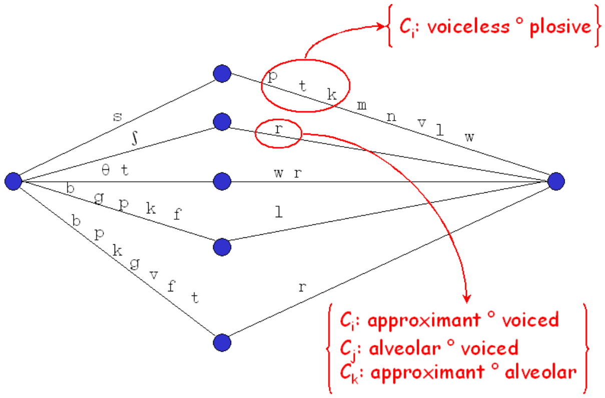

C_i: voiceless ° plosive
p t k
m n v l w
s
ʃ
r
θ t
w r
b g p k f
l
b p k g v f t
r
C_i: approximant ° voiced
C_j: alveolar ° voiced
C_k: approximant ° alveolar

Figure 1: Subsection of phonotactic automaton for English CC- onsets

The multilinear representation consists of phonological features which have been constructed based on acoustic features extracted from the speech signal; each feature has a start and end point in terms of signal time in milliseconds.

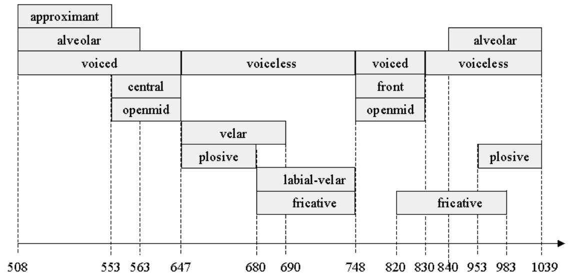


Figure 2: A multilinear representation of the utterance *request*

Phonological parsing in LIPS is guided by the phonotactic automaton which provides top-down constraints on the interpretation of the multilinear representation, specifying which overlap and precedence relations are expected by the phonotactics. If the constraints are satisfied, the parser moves on to the next state in the automaton. Each time a final state of the automaton is reached, a well-formed syllable has been found which is then passed to a corpus lexicon. The lexicon distinguishes between actual and potential syllables (cf. figure 3).

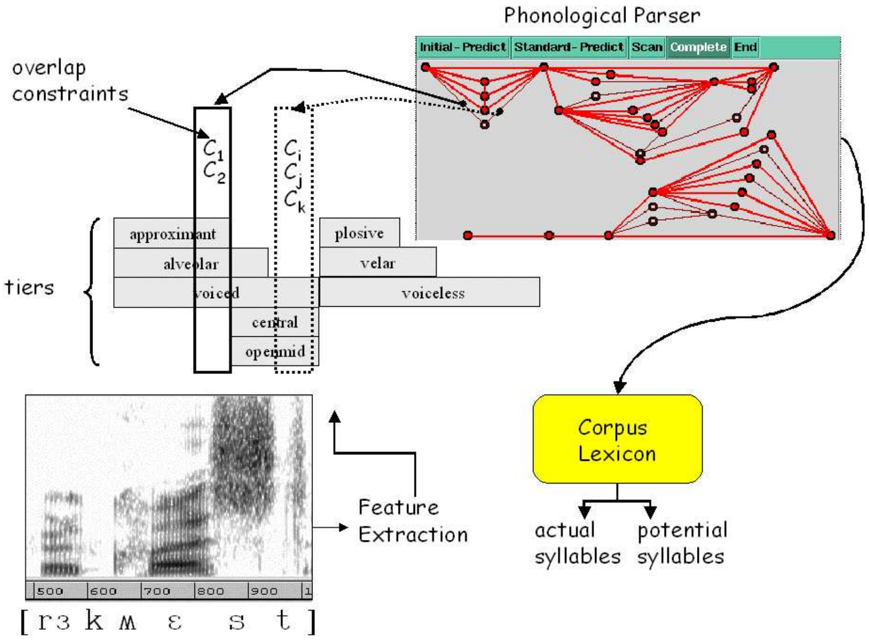


Figure 3: Architecture of the *Time Map* model

Since input to the phonological parser is, in general underspecified due to noise, the *Time Map* model must provide a means of minimising the discrepancy between the expectations defined in the top-down constraints and the actual data by allowing constraint relaxation and output extrapolation. These issues are the topic of the next section.

3 Constraint Relaxation and Output Extrapolation

Carson-Berndsen (2000) identified areas in which statistical information can play a role at different levels of granularity within the *Time Map* model and discussed these with respect to the overall architecture (cf. figure 3). The first area of integration concerns *constraint ranking* which is the lowest level of granularity in that the constraints refer to individual temporal relations. The second integration area is in connection with the *transition weighting* of the phonotactic automaton. Transition weighting is a higher level of granularity in that the whole transition is weighted rather than individual constraints. The third integration area for statistics is the lexicon that refers to a yet higher level of granularity, namely the syllable. This paper goes a stage further in that the first two levels of granularity, namely constraint ranking and transition weighting, are addressed more fully and a means of incorporating them into the model is proposed. We do not comment any further on lexicon issues here, however.

The notion of constraint ranking plays an important role in constraint relaxation and output extrapolation. Constraint ranking can be based on a number of factors: linguistic-preferential, cognitive and statistical. Linguistic-preferential refers to issues of markedness and defaults, cognitive refers to human processing issues and statistical refers to data-oriented issues. Carson-Berndsen (2000) concentrated primarily on corpus-based ranking, but did state that it is more likely that a combination of these factors will be most appropriate for constraint ranking, and for this reason LIPS allows parameters to be chosen and manipulated in order to find the optimal balance between maximal recognition rates and minimal analysis overhead (Carson-Berndsen and Walsh 2000b).

Constraint relaxation should be performed in the model if only *some* of the constraints specified by the phonotactic automaton can be satisfied. As it stands, this is a very arbitrary statement. However, when coupled with a constraint ranking, it becomes a method for dealing with variability and underspecification in the input representation. Constraint ranking is a data-oriented ordering of constraints in particular phonotactic contexts. For example, constraints may be ranked with respect to frequency, duration and percentage overlap of features in specific structural contexts. This information can either be specific to a single corpus or may be based on data from several different corpora. Based on this ranking, constraint relaxation can be applied when an infrequent feature is encountered or a duration is outside a given standard deviation. Furthermore, it is possible to combine this type of ranking with cognitive factors in order to go beyond a corpus-dependent ordering (Carson-Berndsen and Joue 2000). This approach will be discussed further in section 5 below. Constraint relaxation can then be regarded as a means by which particular constraints on an input representation can be ignored. Output extrapolation, on the other hand, is performed to further specify the output representation if the constraints specify expectations that do not conflict with information found in the input. The application of output extrapolation does not guarantee that the output syllable structures are fully specified, however, only that they are well-formed. Here again, a ranking of the constraints, which can participate in output extrapolation, is required.

In LIPS, we distinguish between online processing where speech utterances are interpreted using the constraints and constraint rankings, and offline processing, which is concerned with finding the optimal parameters and constraint rankings for the system (cf. figure 4). In what follows, the integration of constraint rankings into the model are discussed. While the constraint rankings refer to the lowest level of granularity, i.e. individually ranked constraints on temporal overlap relations between features, taken collectively these rankings also provide the basis for the weighting in the phonotactic automaton through the use of transition thresholds, i.e. the next level of granularity.

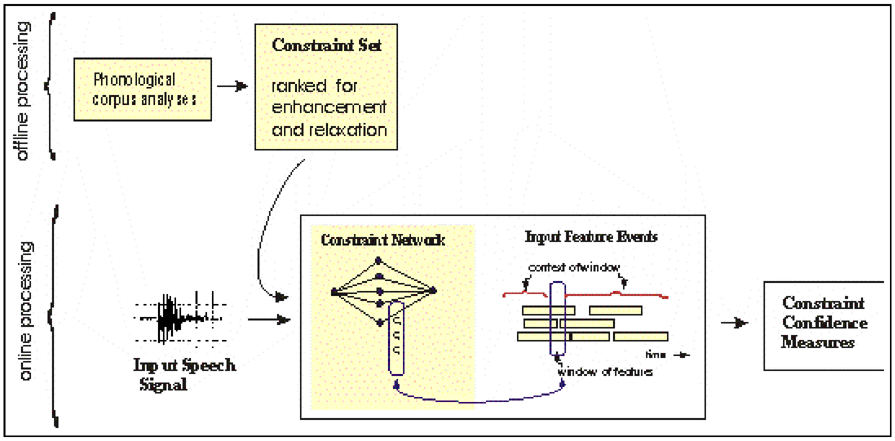


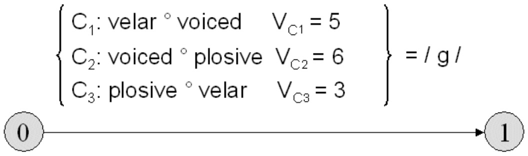
Figure 4: Online and Offline Processing with LIPS

4 Incorporating Constraint Ranking into LIPS

Constraint ranking is incorporated into the phonotactic automaton during the offline processing stage. A transition in the phonotactic automaton may have a number of constraints c_1, c_2, c_3, ... Each of these are constraints on overlap and precedence relations between features that are to be satisfied. For example, c_1 can specify that feature f_1 overlaps feature f_2 ; this is represented as follows: [2]

(1) $c_1 = f_1 \circ f_2$.

Each constraint has a ranking value i.e. c_1 has a ranking value v_{c_1}. Thus, a transition that involves constraints on overlap relations representing a /g/ may look like this:


C1: velar ° voiced V C1 = 5
C2: voiced ° plosive V C2 = 6
C3: plosive ° velar V C3 = 3
= / g /
0
1

Here the second constraint is ranked highest (i.e. largest v) and the third constraint ranked lowest. Rankings of constraints reflect influence of constraints on previous transitions (left-context dependencies) and other effects such as syllable position. In order for a transition to be traversed, the total values of constraints satisfied by the speech input must exceed the threshold on the transition. If a threshold of 9 is assigned to this transition, for example, then c_2 is strengthened

[2] The ∘ symbol indicates the overlap relation

which has the effect of enhancing the constraint: it must now be satisfied in order for the transition to be taken. If the threshold is low, it allows for relaxation of one or more of the constraints on the speech input representation. Adjusting the threshold values so that not all constraints need be satisfied in order to traverse the transition, copes with underspecification in the input representation. The threshold is a parameter of the automaton transition that can depend on speech variables such as rate or register. Ultimately the phonotactic automaton may be able to learn to adjust rankings of constraints and thresholds through training.

Given the input representation of the speech utterance and phonotactic automaton, there may potentially be no best transition from the current node in the automaton. This ambiguity may arise because the input speech does not satisfy enough constraints on all possible transitions from the current node to provide enough weight to traverse any of the transitions. If the speech input does not satisfy *any* constraints on the possible current transitions, then although parsing fails, the diagnostic evaluation procedure of LIPS allows partial analyses which indicate whether output extrapolation at the level of the transition should be undertaken . If some constraints are satisfied, a further constraint relaxation is done by considering right context constraints.

Right-context information can either contribute as immediate *transition resolvers* or to *diagnostically rank* the hypothesis space of the phonotactic automaton. In either case, right-context dependency in the proposed model requires a set of *constraint variation tendencies* between the sets of constraints on each possible transition pair.[3] In a sense, these tendencies indicate what could be missing in underspecified speech input or what processes could have occurred in the speech input to cause a change from the intended speech and are used for output extrapolation. Weights associated with each constraint variation tendency allows different constraints to be relaxed to different degrees. Although the weights on these tendencies can also be used collectively to adjust the threshold of the transitions in question, we favour adjusting values on individual constraints as it provides finer tuning and finer distinctions of the influence of different constraints.

Each constraint variation tendency has the basic form

(2) $pot(C_i) \prec pot(C_j) \Rightarrow C_k \prec C_l, w$

where

- $pot(C_i)$ is a potential constraint (or set of constraints) on a given transition,
- $pot(C_j)$ is a potential constraint (or set of constraints) on a following transition,
- C_k is a constraint or a set of constraints on the given transition that was actually satisfied by the speech input,
- C_l is a constraint or a set of constraints on a following transition that is satisfied by the speech input, and

[3] For now, right context is considered as the next possible single transition from the current transition. Thus the constraint variation tendencies relate the constraints on pairs of transitions.

- w is the probability that the given constraint variation tendency holds.

The tendency can be read in several ways: when the satisfied constraints C_k on transition t_1 precede the satisfied constraints C_l on a following transition t_2, then there is a w probability that constraints $pot(C_i)$ on t_1 and constraints $pot(C_j)$ should be relaxed. Similarly, it can be read as: When the language specifies that feature overlap relations of $pot(C_i)$ should precede $pot(C_j)$, there is a likelihood of w that only the feature events C_k preceding C_l will be seen in the real speech input.

In order for a precedence relation to apply, the speech input must satisfy the implication of the relation ($C_k \prec C_l$) and the condition ($pot(C_i) \prec pot(C_j)$) must include constraints that exist on the transitions of the phonotactic automaton in question. If such is the case, the constraints in the condition of the relations are hypothesised to be potentially present in the speech input and extrapolated in the output, but its ranking value on the transition in the automaton is scaled by the percentage indicated by w. In other words, if there is a value w for a given sequence of satisfied constraints $C_k \prec C_l$ and a sequence of possibly satisfied constraints $pot(C_i) \prec pot(C_j)$, then w scales the ranking value(s) of C_i to account for the possibility that a subset of constraints in C_i does not occur because of constraints C_l. Not all constraints on transition pairs need to be involved in precedence relations, and such a relation can exist for any combination of sequences of constraints.

Once all constraint ranking adjustments are completed using relevant precedence relations, the adjusted constraint rankings are totaled for each transition. The best transition is the one with the greatest scaled positive total distance from the threshold:

(3) $d_t = (\Sigma v_t - \theta_t)/\theta_t$

If d for each candidate transition is the same, then the transition with the highest scaled total of actually satisfied constraints will be considered the better transition. That is, the best transition would have the highest $d_{t_{act}} = (\Sigma v_{t_{act}} - \theta_t)/\theta_t$, where each $v_{t_{act}}$ is only a value of constraints explicitly satisfied by the speech input with no adjustments from the constraint variation tendencies list.

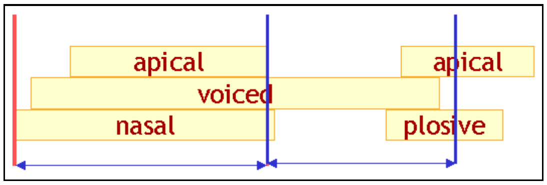
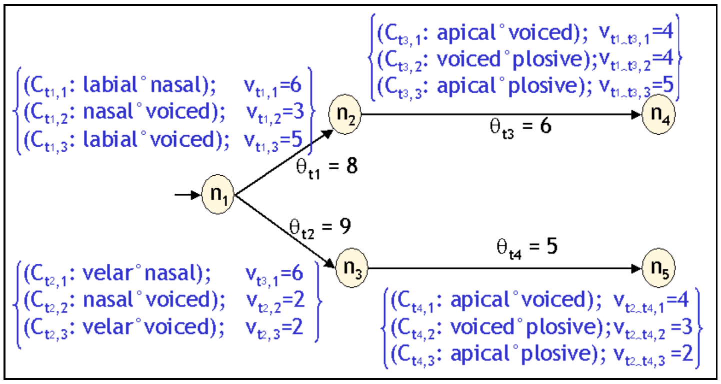


Figure 5: Example multilinear input representation



(C$_{t1,1}$: labial°nasal); v$_{t1,1}$=6
(C$_{t1,2}$: nasal°voiced); v$_{t1,2}$=3
(C$_{t1,3}$: labial°voiced); v$_{t1,3}$=5
(C$_{t3,1}$: apical°voiced); v$_{t1\frown t3,1}$=4
(C$_{t3,2}$: voiced°plosive);v$_{t1\frown t3,2}$=4
(C$_{t3,3}$: apical°plosive); v$_{t1\frown t3,3}$=5
n$_1$
n$_2$
n$_3$
n$_4$
n$_5$
θ$_{t1}$ = 8
θ$_{t2}$ = 9
θ$_{t3}$ = 6
θ$_{t4}$ = 5
(C$_{t2,1}$: velar°nasal); v$_{t3,1}$=6
(C$_{t2,2}$: nasal°voiced); v$_{t2,2}$=2
(C$_{t2,3}$: velar°voiced); v$_{t2,3}$=2
(C$_{t4,1}$: apical°voiced); v$_{t2\frown t4,1}$=4
(C$_{t4,2}$: voiced°plosive);v$_{t2\frown t4,2}$ =3
(C$_{t4,3}$: apical°plosive); v$_{t2\frown t4,3}$ =2

Figure 6: Example transitions of phonotactic automaton

To illustrate this constraint ranking method for relaxation and enhancement, suppose we have a multilinear speech input representation as in figure 5, a phonotactic automaton as in figure 6, and a set of constraint variation tendencies:

(4) labial ∘ nasal ≺ apical ∘ plosive ⇒ apical ∘ nasal ≺ apical ∘ plosive, w =20%

(5) velar ∘ voiced ≺ apical ∘ voiced ⇒ apical ∘ voiced ≺ apical ∘ voiced, w =10%

As shown in figure 5, the rankings of constraints on transition t_3 depend on the constraints on transition t_1 i.e. $v_{t_1 \frown t_3}$ (likewise for transition t_4 on t_2), so even though the exact constraints on t_3 occur also on t_4, the different respective preceding constraints (on t_1 and t_2) lead to different ranking values of the constraints.

The threshold on transition t_1 is θ_{t_1} = 8 (to enhance constraints $c_{t_1,1}$, or $c_{t_1,3}$ in conjunction with at least another constraint) and the threshold on transition t_2 is θ_{t_2} = 9, but our input speech stream satisfies only $c_{t_1,2}$ and $c_{t_2,2}$. Since the speech stream is too ambiguous, the model looks ahead to the next possible transitions (t_3 and t_4) for possibilities of constraint relaxation on transitions t_1 or t_2. It checks which constraints on t_3 and t_4 are satisfied and compares all the satisfied constraints to a list of constraint variation tendencies (in our example, there are only two tendencies as given in variation 4 and variation 5).

A constraint variation tendency can only be applied if the constraints to the right of the implication are satisfied by the speech input and if the conditions of the tendency are constraints that actually occur in the phonotactic automaton. Although the the implication of variation 5 (apical ∘ voiced ≺ apical ∘ voiced) occurs as constraints on transition t_3 and are constraints satisfied by the speech input in Figure 5, the first part of the condition of variation 5 (velar ∘ voiced) does not occur as constraints on transition t_1. Thus, constraint variation tendency 5 does not apply at all. Constraint variation tendency 4, however, applies to the transition pair $t_1 \frown t_3$ as the implication of variation 4 is satisfied by the speech input in the current and next windows and its condition covers constraints which occur on transitions t_1 and t_3 of the automaton.

The tendency formulated in variation 4 informs us that we can relax constraints $c_{t_1,1}$ and $c_{t_1,3}$. To indicate that these constraints were relaxed, the ranking values of these constraints can either be scaled by w of variation 4, or the threshold of the transition (t_1) be scaled. We chose to adjust ranking values instead as this allows finer tuning of constraint ranking. Thus, the new value for $c_{t_1,1}$ is 1.2 (= $w * v_{t_1,1}$ = 0.2 * 6) and for $c_{t_1,3}$ is 1.0 (= 0.2 * 5). With these new values, the total value from satisfied and possibly satisfied constraints is 5.2 (=1.2+3+1), which does not clear the threshold of 8 on transition t_1. If variation 4 were applicable to the transition pair $t_2 \frown t_4$, adjusting ranking values should also be done.

Transition t_1 is a better transition than t_2, as its scaled distance from the threshold is closer. The path ultimately traversed from t_1 onwards might not lead to the best syllable hypothesis, however. The sum of scaled distances from the transition threshold (Equation (3)), provides a method of ranking transitions, and ultimately paths, through the automaton. It offers a diagnostic evaluation of the automaton's hypotheses. As a diagnostic tool, it does not force a decision within n-arcs time, but the decision is made based on comparison of the confidence values for the hypotheses for the entire path through the automaton.

5 A Functional Cognitive Basis for Constraint Rankings

So far this paper has generally discussed how constraint relaxation and output extrapolation can be incorporated into the *Time Map* model to improve robustness. However, we have avoided to specify the exact nature of how ranking of the constraints is achieved. There are doubtless many strategies, but we argue for a functional cognitive paradigm for ranking constraints. This paradigm is based on Phonology as Human Behavior (PHB), a combinatorial phonological analysis which argues that the skewed distribution of speech sounds are structured because of a collaborative relationship between human articulatory constraints and perceptual constraints for efficient communication (Tobin 1997, Diver 1995). Thus, to explain the nonrandom structure in speech sounds, PHB focuses on identifying the constraints and the interactions among them, and identifying the features of speech sounds which are involved in these interactions.

PHB posits features which are indicative of gestural control and coordination including, for example, active articulators (*apex*, *velum*, *larynx*), the type of gestural movement (*mobile*, *stable*), or number of articulators involved in the production of a given speech sound. The theory describes the interdependency of these features with perceptual constraints for the goal of communication in terms of interacting disfavourings or tendencies. For example, PHB posits the disfavouring for the same active articulator to be used in adjacent sounds, as in *tl*, and the favouring of vowels (easy to articulate) but the disfavouring of too many vowels (difficult to make perceptual distinctions).

Understanding these direct factors that shape speech sounds and the interactions of such factors, leads to motivated predictions of the dynamic structural tendencies or changes in speech. Moreover, since in speech the constraints are based on human physiological considerations, the set of constraints should be similar across languages. The interaction of constraints with each other (the ranking of constraints) is then language-specific and even speaker- and context-specific.

The ranking value v of each constraint in our model is based on corpus analyses of pairwise distributions of temporal relations between features, i.e., how likely the sequence of a given set of overlap relations preceding another set occurs. The higher the value of v, the more frequently sounds with these features occur, and satisfying the constraints of these events have greater relative weight and are less likely to be relaxed. The lower the value of v, the less frequent the sequence of events occurs and thus the more frequent these constraints should be relaxed.

The ranking value for a constraint can also be dependent on any number of constraints in the preceding transition. For example, the ranking value of $c_{t_2,1}$ may depend on the fact that it follows $c_{t_1,1} \circ c_{t_1,3}$. The ranking value is not the actual distribution of the relevant feature relations in the corpus, but is scaled by the total percentage of all the constraints on the transition pairs. So if according to the corpus the distribution of $c_{t_1,1} \circ c_{t_1,3} \prec c_{t_2,1}$ is 20% then the ranking value for $c_{t_2,1}$ is 20/(20 + $v_{c_{t_2},2}$ + $v_{c_{t_2},3}$ + . . .). This allows a scaled evaluation of how likely transition t_2 is to be traversed.

In this cognitive approach, the constraint variation tendency list is compiled according to favourings/disfavourings that constrain the development and production of speech as defined by PHB. The constraint variation tendencies then indicate the type of variations on the automaton constraints that may occur due to human physiological and behavioural factors.

6 Conclusion

This paper has been concerned with constraint relaxation and output extrapolation procedures in a computational linguistic model for speech recognition. In this model, a constraint ranking provides the basis for initiating these procedures for robust interpretation of multilinear representations of speech utterances. The generic development environment for the computational linguistic model has both an online and an offline functionality which allows optimum incorporation of statistical information to be further investigated. The development environment has been specifically designed to extend to phonotactic descriptions of other languages allowing, on the one hand, specific constraint rankings to be integrated and, on the other hand, facilitating the investigation of more language independent constraints based on cognitive factors such as those suggested by Phonology as Human Behavior. Current work now involves diagnostically evaluating phonotactic descriptions of other languages in the context of speech recognition and it is anticipated that this will provide insights into the choice of feature sets which are optimal to the task.

References

Carson-Berndsen, J.(1998), *Time–Map Phonology: Finite State Models and Event Logics in Speech Recognition*, Kluwer Academic Publishers.

Carson-Berndsen, J.(2000), Computers, language and speech: Integrating formal theories and statistical data, **358**(1770), 1255–1266.

Carson-Berndsen, J. and Joue, G.(2000), Cognitive constraints in a computational linguistic model for speech recognition, *Proceedings of AICS, Galway, August 2000*.

Carson-Berndsen, J. and Walsh, M.(2000a), Generic techniques for multilingual speech applications, *Proceedings of TALN 2000, Lausanne, October 2000*.

Carson-Berndsen, J. and Walsh, M.(2000b), Interpreting multilinear representations of speech, *in* M. Barlow (ed.), *Proceedings of the Eighth Australian International Conference on Speech Science and Technology, Canberra, December 2000*, pp. 472–477.

Diver, W.(1995), The theory, *in* E. Contini-Morava and B. Goldberg (eds), *Meaning as Explanation: Advances in Sign-Oriented Linguistic Theory*, Mouton de Gruyter, Berlin and New York, pp. 45–13.

Tobin, Y.(1997), *Phonology as Human Behavior: Theoretical Implications and Clinical Applications*, Duke University Press.