Defining Constraints for Multilinear Speech Processing*

Julie Carson-Berndsen and Michael Walsh

Department of Computer Science
University College Dublin
`{ julie.berndsen, michael.j.walsh }@ucd.ie`

Abstract

This paper presents a constraint model for the interpretation of multilinear representations of speech utterances which can provide important fine-grained information for speech recognition applications. The model uses explicit structural constraints specifying overlap and precedence relations between features in both the phonological and the phonetic domains in order to recognise well-formed syllable structures. In the phonological domain, these constraints together form a complete phonotactic description of the language, while in the phonetic domain, the constraints define the internal structure of phonologial features based on phonetic realisations. The constraints are enhanced by a constraint relaxation procedure to cater for underspecified input and allows output representations to be extrapolated based on the phonetic and phonological information contained in the constraints and the rankings which have been assigned to them. This approach thus addresses issues of robustness in speech recognition.

1. Introduction

This paper presents a computational linguistic model which has been developed for the explicit purpose of providing fine-grained structural information for speech technology applications. The model has been described in detail elsewhere [1,2] but we review the model below with explicit reference to the types of constraints it assumes and discuss how these have been enhanced to address the notion of robustness in speech recognition. Our concern in this paper is not to compare the performance of our model with current stochastic approaches to speech recognition, but to highlight areas in which we believe explicit knowledge constraints can contribute to speaker- and corpus-independent speech recognition and reduce the need for training data. This is very much in line with parallel research by Deng [3,4] who proposed an autosegmental feature-based approach to generating word pronunciation models represented as finite state automata which were then interfaced with a trended HMM. While our motivation is very similar, the constraint-based model is based on a complete phonotactic description of a language which is used to provide top-down constraints on such multilinear (autosegmental) feature representations.

The next section briefly reviews the constraint-based model in the context of speech recognition. Section 3 discusses the application of phonotactic constraints in the phonological domain and section 4 demonstrates how this technique can be extended to the phonetic domain to allow for a representation which is more closely related to the speech signal. Section 5 describes how the constraints in each domain can be relaxed in order to extrapolate the output representations given noisy input data. Section 6 concludes with some discussion of future work.

2. The Time Map Model

The *Time Map* model was proposed as a computational linguistic model for speech recognition by Carson-Berndsen [1] 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 [5]. 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.

The *Time Map* model uses a finite-state network representation of the phonotactic constraints in a language, known as a phonotactic automaton (cf. section 3 below), together with axioms of event logic to interpret multilinear representations of speech utterances. In order to provide speech utterances with a phonological interpretation, this approach encompasses both an absolute time level, in which speech signals are related to speech events by temporal annotations (in terms of millisecond values), and a relative time level, in which speech events are related to each other in terms of overlap and precedence relations. The architecture of the model in the context of speech recognition is depicted in figure 1.

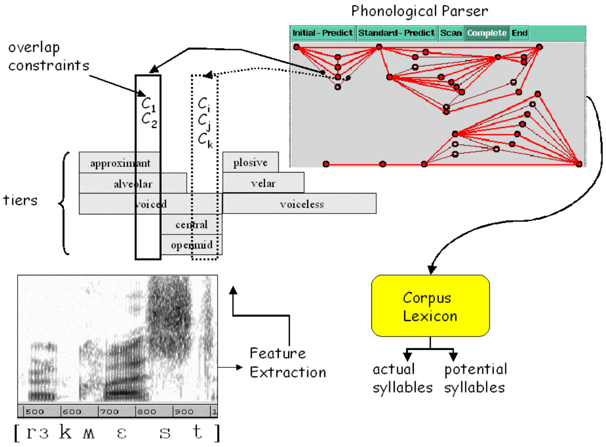

Figure 1: Time Map Architecture

Input to the model is a multilinear representation of a speech utterance in terms of absolute time events. i.e. features with start and end points which are extracted from the speech signal (we will elaborate on this in section 4 below). Phonological parsing in the *Time Map* model 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 passed then to a corpus lexicon which distinguishes between actual and potential syllables. The corpus lexicon is compiled from the generic lexicon described in [6].

3. Phonotactic Constraints

The primary knowledge component of the model is a complete set of phonotactic constraints for a language which is represented in terms of a finite state automaton. A subsection of a phonotactic automaton for CC- combinations in English syllable onsets can be seen in figure 2.

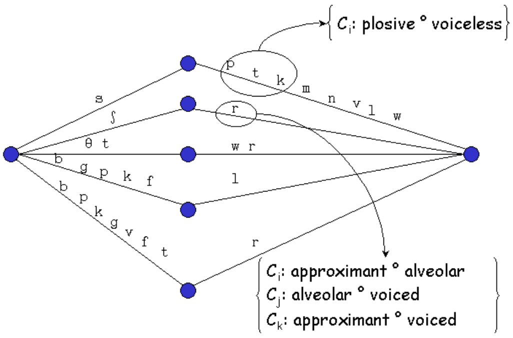
C_i: plosive ° voiceless
s
ʃ
θ t
b
g p k f
b p k g v f t
p t k m n v l w
r
w r
l
r
C_i: approximant ° alveolar
C_j: alveolar ° voiced
C_k: approximant ° voiced

Figure 2: English CC- onsets

The arcs 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] The phonological features used in this representation are multi-valued IPA-like features which are defined with respect to a tier model where the tiers define phonation, manner of articulation, place of articulation, vowel height, rounding and tenseness. The constraint *Ci: plosive ° voiceless*, for example, states that the feature *plosive* on the manner tier should overlap the feature *voiceless* on the phonation tier. The multilinear input representation which is constrained by this phonotactic automaton, in this case, is a tiered representation of such phonological features analogous to an autosegmental score [7] and not unlike that used in the synthesis model of articulatory phonology [8]. There is, therefore, no strict segmentation of the input at the level of the phone or phoneme and the constraints apply not to the actual temporal annotations of the input but to the temporal relations which exist between them.

The advantage of the phonotactic constraints is that they restrict all outputs of the model to structures which are well-formed in the language and are, therefore, a means of treating out-of-vocabulary items. Partial analyses may also be output for diagnostic purposes and extrapolation. While this addresses the notion of robustness in speech recognition to a certain extent, the main criticism we would have of our original model, is that it did not take any statistical knowledge into account and thus provided no means of ranking the output hypotheses. For this reason, we have extended the phonotactic automaton and the parsing procedure to incorporate a number of additional constraints.

Firstly, each feature participating in a constraint is now augmented by an average duration parameter with respect to the particular phonotactic context in which it appears. These average durations are calculated on the basis of a large body of data, but are not intended to be corpus- or speaker-specific and may need to be tuned to reflect speech rate. The duration parameter is used to define the prediction space for the next arc in the phonotactic automaton during processing.

Secondly, we have integrated a constraint ranking methodology into the model by allowing constraints on the arcs to be ranked and an overall threshold to be defined which provides the basis for constraint relaxation. The aim of constraint relaxation is to cater for imperfect and noisy input by allowing constraints to be relaxed and output representations to be extrapolated based on structural information defined in the phonotactics. This will be discussed in more detail in section 5.

Despite the completeness of the phonotactic constraints and the fact that they are not dependent on segmentation of the input into non-overlapping phonemic units, the model has in the past been viewed as a primarily phonological approach based on features which are not apparent in the signal. We now address this issue in section 4.

4. Towards Phonetic Constraints

The LIPS generic development environment for the *Time Map* model is independent of any particular feature set and allows users to define the phonotactic automaton with respect to any feature set. The features in the input representation are treated autonomously, however. In this context, we distinguish between two different types of event within a knowledge domain: simplex events and complex events. A simplex event has no internal structure with respect to a knowledge domain. For example, the phonotactic automaton in section 3 assumed that the feature *plosive* is a simplex feature in the phonological domain. It becomes a simplex event when coupled with a temporal annotation. A syllable is a complex event in the phonological domain, however, as it has an internal structure based on the composition of the simplex events.

This notion can be extended to the phonetic domain (cf. [1]). We assume that, at the phonetic level, a feature such as *plosive* is indeed complex consisting of combinations of articulatory movements or acoustic manifestations. A complex *plosive* feature may consist of simplex events such as *closure*, *release*, *frication*, which will be realised differently depending on the context in which it occurs. In English, a voiceless plosive in syllable initial position before a vowel will realise all three of these features. However, in syllable

[1] The monadic symbols written on the arcs in figure 2 are purely mnemonic for the feature overlap constraints they represent.

final position, the *release* may not be apparent at all. Church [9] emphasised the importance of such allophonic information in providing cues for parsing in speech recognition. Furthermore, the information on the transition phase from *plosive* to *vowel* should also be included as this provides useful indications. An increase in F_1 frequency during the transition to the vowel is characteristic of all plosive-vowel transitions and a cue to manner of production [10].

In order to cope with this type of predictable variation in the model, we extend the phonotactic automaton by a transduction relation which maps between phonological feature automata and a set of constraints on their overlap relations. [2] A phonological feature automaton defines the internal structure of a complex event in the phonetic domain. The transduction relation defines the many-to-one mapping between the phonological feature automaton and the phonological feature itself. An example of a possible *plosive* phonological feature automation in initial syllable position before a vowel is depicted in figure 3.

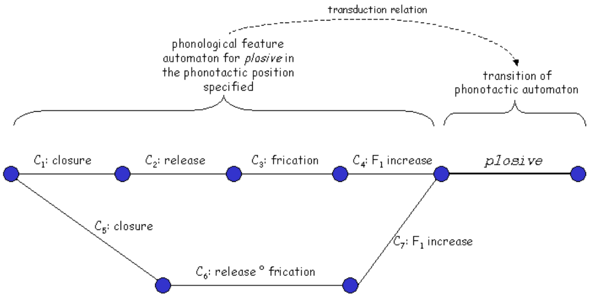
transduction relation
phonological feature automaton for *plosive* in the phonotactic position specified
transition of phonotactic automaton
C_1: closure
C_2: release
C_3: frication
C_4: F_1 increase
plosive
C_5: closure
C_6: release ° frication
C_7: F_1 increase

Figure 3: An example phonological feature automaton for *plosive*

Similarly, the place of articulation feature, for the plosive in this phonotactic context before a vowel, will be heavily influenced by the formant transitions of the following vowel. Thus, a phonological feature for the place feature of the plosive would expect to use a constraint on the F_2 and possibly F_3 transitions which seem to be sensitive to the place of articulation. In particular, the starting frequency (or locus) of F_2 may be used.

The set of phonological feature automata should be constructed with respect to phonetic or acoustic features which can be detected in the signal and may come from a combination of techniques rather than just assuming that these are extracted by HMMs. Clearly the phonological feature automata and the phonotactic automaton are sets of top-down constraints which we do not expect to be fulfilled in all cases in natural speech. For this reason we assume that constraint relaxation can be applied in both the phonetic and the phonological domains. This is described in the next section.

5. Constraint Relaxation

Constraint relaxation should be performed in the model if only *some* of the constraints specified by either the phonotactic or phonological feature automata 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 [12]. Constraint relaxation can then be regarded as a means by which particular constraints on an input representation can be ignored. We illustrate constraint relaxation simply using the following three constraints on overlap relations on a particular transition of a phonotactic automaton:

- *C_1: plosive ° voiced* with ranking 6
- *C_2: voiced ° labial* with ranking 4
- *C_3: plosive ° labial* with ranking 3

Assuming the transition has a threshold in this phonotactic context of 7, then at least two of these constraints must be satisfied (i.e. the sum of the rankings must exceed the threshold) in order for this transition to be taken.

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. Here again, a ranking of the constraints, which can participate in output extrapolation, is required. We use same constraints *C_1*, *C_2*, *C_3*, for illustration but this time we are referring to another transition which has a threshold value of 6. Given an input representation which is underspecified with respect to place of articulation, then constraint relaxation will allow the transition to be taken with only *C_1* satisfied. Output extrapolation can then be performed using *C_2* and *C_3* if there is no information in the input representation which explicitly conflicts with these constraints such as another place of articulation, for example. So although the input representation was underspecified with respect to place of articulation, the output representation will contain this information as it was augmented using the structural knowledge contained in the phonotactic automaton. The application of output extrapolation does not guarantee that the output syllable structures are fully specified, however, only that they are well-formed.

In LIPS, a distinction is made 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. While the constraint rankings refer to individually ranked constraints on temporal overlap relations between phonological or phonetic features, taken collectively these rankings also provide the basis for weighting in the

[2] This is not unrelated to the allophonic parser described in [11] except that the parser defined there assumed a segmentation of the input into a string of allophones.

phonotactic automaton and the phonological feature automata respectively, through the use of transition thresholds.

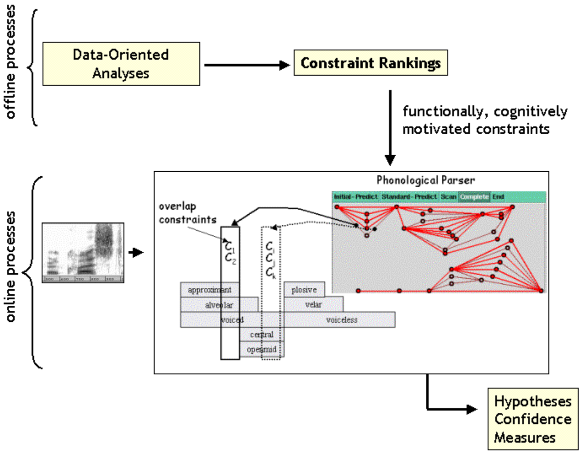

Figure 4: Offline vs. Online Processing in LIPS

It goes beyond the bounds of this paper to describe the constraint relaxation and output extrapolation procedures in further detail. Further information can be found in [13] and in a paper to appear shortly.

6. Conclusion

This paper has presented a constraint model for interpreting multilinear representations of speech utterances which can provide important fine-grained information for speech recognition applications. It was demonstrated that the model integrates phonotactic and phonetic information in a nonsegmental fashion. Currently, we are working on developing a wider range of phonological feature automata and investigating how they contribute to robustness in the model through the use of the constraint relaxation and output extrapolation techniques which were described above. While our ongoing research is directed towards optimising this model through the use of statistical information and cognitive constraint rankings, the model also provides useful constraints for fine-tuning more stochastic approaches for robust speech applications. Future directions for this work are to investigate the application of the computational model in the context of speech synthesis whereby multilinear phonetic representations are generated rather than interpreted by the model and serve as parameters to a synthesis module. We believe that the use of the same model for both recognition and synthesis will provide insights into the different levels of granularity of information required for truly robust speech applications and lead to approaches which combine linguistic and statistical knowledge more explicitly. Furthermore, the integration of phonotactic descriptions of other languages adds a dimension of multilinguality to our model which will support the incorporation of phonetic features which are common among languages.

7. References

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

[2] Carson-Berndsen, J., "Finite State Models, Event Logics and Statistics in Speech Recognition", In Gazdar, G.; K. Sparck Jones & R. Needham (eds.): *Computers, Language and Speech: Integrating formal theories and statistical data*. Philosophical Transactions of the Royal Society, Series A, 358(1770), 1255-1266, 2000.

[3] Deng, L. "Speech Recognition Using Autosegmental Representation of Phonological Units with Interface to the Trended HMM". *Free Speech Journal*, 1997.

[4] Deng, L. "A dynamic feature-based approach to the interface between phonology and phonetics for speech modelling and recognition." *Speech Communication*, 24, 299-323, 1998.

[5] Carson-Berndsen, J. and Walsh, M., "Generic techniques for multilingual speech technology applications", *Proceedings of the 7th Conference on Automatic Natural Language Processing*, Lausanne, Switzerland, 61-70, 2000.

[6] Carson-Berndsen, J. "A Generic Lexicon Tool for Word Model Definition in Multimodal Applications". *Proceedings of EUROSPEECH 99*, 6th European Conference on Speech Communication and Technology, Budapest, September 1999

[7] Goldsmith, J. *Autosegmental and Metrical Phonology*. Basil Blackwell, Cambridge, MA. 1990.

[8] Browman, C.P. and Goldstein, L., "Articulatory gestures as phonological units", *Phonology 6*, Cambridge University Press, Cambridge, 201-251, 1989.

[9] Church, K. W. *Phonological Parsing in Speech Recognition*. Kluwer Academic Publishers, Boston. 1987.

[10] Kent, R.D. and C. Read. *The Acoustic Analysis of Speech*. Whurr Publishers, 1992.

[11] Carson. J. "Unification and Transduction in Computational Phonology In: *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, vol 1, 106-111, 1988.

[12] Carson-Berndsen, J. and Joue, G., "Cognitive constraints in a computational linguistic model for speech recognition", *Proceedings of the 11th Irish Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland, 2000.

[13] Carson-Berndsen, J. and Walsh, M., "Interpreting Multilinear Representations in Speech". *Proceedings of the 8th Australian Conference on Speech Science and Technology*,Canberra, Australia, 472-477, 2000.

* The authors gratefully acknowledge support from Enterprise Ireland under grant SC/99/069.