PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL PHYSICAL & ENGINEERING SCIENCES

Finite-state models, event logics and statistics in speech recognition

Julie Carson -Berndsen

Phil. Trans. R. Soc. Lond. A 2000 **358**, 1255-1266
doi: 10.1098/rsta.2000.0584

THE ROYAL SOCIETY

Finite-state models, event logics and statistics in speech recognition

By Julie Carson-Berndsen

Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland

This paper presents a constraint-based approach to speech recognition which combines aspects of event logic with efficient processing strategies. Although stochastic approaches are currently at the forefront of speech-recognition applications, it has now been recognized that linguistic structure is required in order to deal with the problem of recognizing new words. The computational linguistic approach presented here offers solutions to the problems of how to process words which have not been heard before, and how to develop fine-grained knowledge representation and processing techniques for linguistic units smaller than the word. Furthermore, it is investigated how statistical data can be integrated into the phonological constraint model in order to minimize the discrepancy between expectations defined in the top-down constraints and the actual data.

Keywords: finite-state techniques; multilinearity; events; speech recognition

1. Introduction

One of the major problems in the area of speech technology concerns the treatment of new words. In general, a new word refers to any structure that is well-formed with respect to the phonological and morphological constraints of a particular language but which is not part of a lexicon of that language. For example, the word *blant* is not found in any English lexicon, but a native speaker of the language would consider it to be well-formed (as opposed to a form such as *bnanlt*, which is considered ill-formed). Such forms point to idiosyncratic gaps in the lexicon and, thus, could potentially become words of the language in the future. In the context of speech recognition, the term 'new word' is usually restricted to mean new with respect to a particular corpus. While treatment of this phenomenon is one of the motivations for the model presented in this paper, the model itself assumes the broader definition of new words to refer to potential forms not included in the lexicon of the language.

In order to recognize or generate new words, information about their internal structure is required. It is now generally accepted that morphological and phonological constraints are required either implicitly (as in hidden Markov models) or explicitly, in speech-technology applications. While there is no doubt that stochastic approaches to speech recognition are at the forefront of current research and form the basis of most commercial applications, Smolensky (1999) has pointed out that 'the potential for phonological theory to improve the performance of speech-recognition systems remains largely unrealized'. This paper describes the *Time Map*

model, a computational model of phonological interpretation for speech-recognition applications developed by the author since the early 1990s and recently reported in Carson-Berndsen (1998).

The *Time Map* model uses finite-state methodology and an event logic to demonstrate how declarative descriptions of phonological constraints can play a role in speech recognition (see also Wagner (1997) for a speech-synthesis application). The main aim of the work has not been to build a speech-recognition system that can compete with stochastic systems in terms of system performance, but, rather, to design a knowledge-based multilinear component for a speech-recognition system that uses phonological well-formedness constraints and which is demonstrably of value in recognizing new words, modelling and investigating coarticulation effects (temporal overlap of properties), and dealing with underspecified structures. The long-term goal of this work is, therefore, to provide a linguistic basis for integrating symbolic and stochastic approaches to speech recognition. While initial research on this model has concentrated purely on the symbolic approach, it has already been demonstrated that the explicit incorporation of phonological knowledge can provide useful structural constraints for the fine tuning of stochastic models (Jusek *et al.* 1994). The penultimate section of this paper investigates the extent to which statistical information can be integrated into the symbolic model for the purposes of application-specific tuning.

The main motivation for the *Time Map* approach to speech recognition concerns the compositionality and variability of spoken language, which can only be catered for to a limited extent by concatenative models, which assume a rigid segmentation into non-overlapping units at some level of granularity (e.g. diphones, phonemes, demi-syllables, syllables). When dealing with speech recognition, inputs are not in the form of discrete non-overlapping elements, as is usual in written language processing, rather they are lattices containing competing hypotheses and possibly conjunctions in the case of overlapping information. The treatment of such gaps and overlaps in the input lattice is required at all levels of processing and is termed the *lattice-to-chart* problem. Furthermore, the predictive skill of the native speaker of a language, which allows the projection of a finite set of actual structures onto a possibly infinite set of potential (well-formed) structures, must also be addressed if the 'new-word' problem is to be solved.

The *Time Map* model of phonological interpretation assumes lattices of acoustic–phonetic events as input and uses parallel finite-state machines together with an event logic to recognize well-formed syllable structures, allowing a distinction to be made between actual (i.e. those in the lexicon) and potential forms. The speech signal is represented as tiers of features which are interpreted in terms of overlap and precedence relations between events, avoiding a rigid segmentation into non-overlapping units and allowing coarticulation variants to be modelled. The constraint-based *Time Map* model of phonological interpretation has been implemented and evaluated within a linguistic speech-recognition system that uses an incremental architecture. Carson-Berndsen (1998) provides further details.

This paper will first introduce the finite-state methodology and motivate the use of multilinear representations of phonological structures in the context of speech recognition. The temporal interpretation of these structures with respect to an event logic will then be presented and the *Time Map* model will be discussed in the context of a linguistic speech-recognition system. Finally, the components of the implemented

model that lend themselves to the incorporation of statistical data will be highlighted in the context of the overall architecture.

2. Finite-state models

Finite-state techniques have been used extensively in both computational phonology and speech recognition. However, it would be fair to say that much of the research into finite-state techniques in phonology has not concentrated particularly on speech-technology applications, and, conversely, statistical finite-state models, such as hidden Markov models, have not, in the majority of cases, made explicit use of phonological constraints. This section will concentrate on those aspects of finite-state phonology that are relevant to the *Time Map* model; further background information can be found in Carson-Berndsen (1998).

Finite-state phonology distinguishes between linear and multilinear representations. Linear finite-state phonology—exhibited in the work of Koskenniemi (1983), Kaplan & Kay (1994) and others—has been very influential in the area of computational linguistics in general, and the methodology has been used in a wide application area covering morphological analysers, spellers, hyphenators, part-of-speech tagging, indexing, and retrieval (see the Web pages of Xerox Research and Lingsoft†). Multilinear finite-state phonology has dealt with the formalization and implementation of nonlinear models such as autosegmental phonology and non-concatenative morphology by Kay (1987), Bird & Ellison (1994), Kornai (1995) and others. The *Time Map* approach is concerned with multilinear finite-state phonology. It differs from the above, however, in that it aims to provide a complete set of constraints on syllable structures and in that it has been integrated into an actual speech-recognition application for German.

Constraints on phonological well-formedness can be represented declaratively in terms of networks which can be interpreted by finite-state automata. Such constraints, known as phonotactics, represent the possible combinations of sounds of a language within a particular phonological domain, usually the syllable. An example of such an automaton representation, of CCV- combinations in German syllable onsets, is depicted in the network of figure 1. The *Time Map* model extends this simple linear model of phonotactic constraints to a multilinear model where labels on the arcs of the automaton representation are no longer simple phonemes, but, rather, represent constraints on temporal overlap relations which occur in each structural position. Examples of such constraints for two transitions are shown in the figure. The arcs specify only those constraints required in the particular structural position, i.e. they are based on natural classes of features and are, in general, underspecified with respect to all the features needed to define any individual sound.

The advantage of this type of representation of phonotactic constraints is that an interpretation of multilinear phonological representations, as found in autosegmental and articulatory phonology, is made possible. Figure 2 shows a multilinear representation of the potential word *blant* including coarticulation. It consists of a set of parallel tiers of features, each of which has its own temporal pattern or melody. Only three phonological tiers are shown for the purposes of illustration. Each tier has its own segmentation and the startpoints and endpoints of each of the features on the tiers may differ, which makes the multilinear representation fundamentally different

† Xerox: http://www.xrce.xerox.com/research/mltt/fst/; Lingsoft: http://www.lingsoft.fi/.

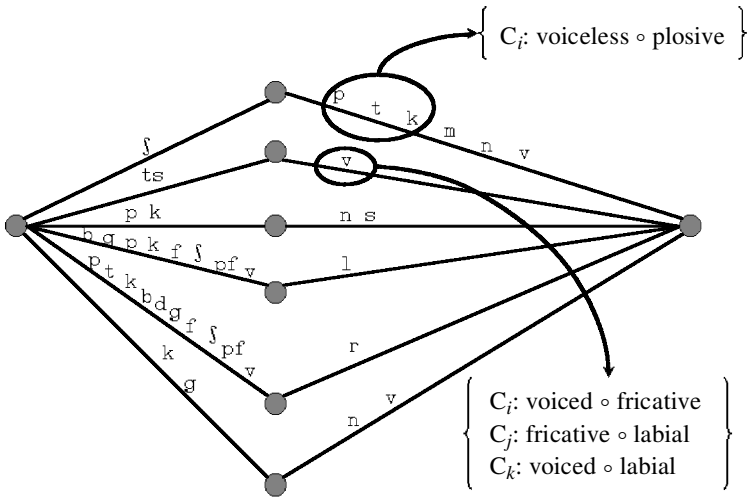

Figure 1. An automaton representation of German CCV- combinations.

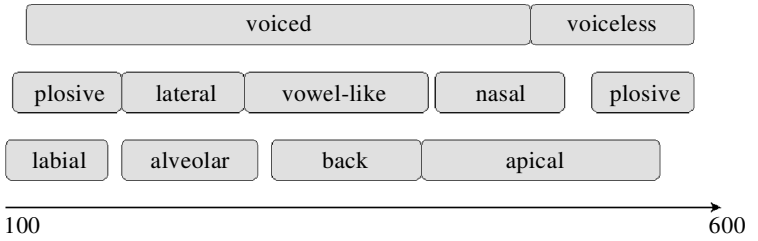

Figure 2. A multilinear representation of the potential word *blant*.

from a standard segmental phonological representation in which all the features are subject to the same segmentation.

The phonotactic automaton defines a structured set of constraints which represents a complete phonotactics of a language and can distinguish between well-formed and ill-formed structures independent of any particular corpus and, indeed, of any particular speaker. The temporal constraints of the phonotactics do not assume that a strict segmentation into non-overlapping units has taken place and, therefore, coarticulation phenomena and many speech variants can be modelled in the multilinear representation.

3. Event logics

The *Time Map* model proposes a flexible non-segmental approach to speech recognition, which incorporates the notion of compositionality by employing several sources of information simultaneously. Such an approach avoids a strict segmentation of the speech signal into phonemes or other units of similar granularity, and, therefore, the acoustic front-end does not have to classify over phonological symbols, but can detect autonomous acoustic features from the signal. The acoustic feature representation is

analogous to the multilinear representation depicted in figure 2, except, of course, that it consists of more than three tiers. This representation of acoustic information in a multilinear structure of sequential and parallel features is nearer to the signal than the phoneme sequences of more traditional phonologies. In order to relate the features in such a representation both to the speech signal and to the temporal constraints specified in the phonotactic automaton, the *Time Map* model uses an event logic based on time-type domains which coexist as different perspectives on spoken-language utterances (see Gibbon 1992). The *Time Map* model is primarily concerned with a *relative-time domain* (T_{rel}), which is an abstract temporal domain in which categories are assumed to have duration and can be viewed as intervals with temporal relations between them, and an *absolute-time domain* (T_{abs}), which is an utterance time domain in which categories have a temporal annotation and, therefore, a direct reference to real signal time.

The problem of representing temporal knowledge and of temporal reasoning with this knowledge has been investigated by Allen (1983), Van Benthem (1983) and Freksa (1992), among others. Building on this work, proposals have been made in computational phonology for providing multilinear phonological representations with a formal interpretation using events and the axioms of event logic (Bird & Klein 1990; Carson-Berndsen 1998), such that feature co-occurrence (or association in autosegmental phonology terms) is interpreted as temporal overlap and sequencing as temporal precedence. Although such an approach allows the explicit introduction of a temporal dimension into the phonological description, a restriction of this interpretation to T_{rel}, as proposed by Bird (1995), has the consequence that no reference can be made to actual speech tokens with absolute-time annotations. Speech recognition, however, requires a mapping from absolute-time annotations to a relative-time domain in which actual time is no longer needed ($\mathrm{T_{abs}} \rightarrow \mathrm{T_{rel}}$).

The *Time Map* model distinguishes between events in T_{rel} and events in T_{abs}.

A relative-time event is defined as an ordered pair $\rho\tau\epsilon = \langle I, F\rangle$, where I refers to some interval in the time domain T_{rel} and F refers to a feature or property of the interval.

For example, in the multilinear representation in figure 2, the feature *voiced* on the phonation tier can be a relative-time event $\rho\tau\epsilon_1 = \langle I_1, \mathrm{voiced_{phonation}}\rangle$.

An absolute-time event is defined as an ordered pair $\alpha\tau\epsilon = \langle\langle t_{\mathrm{s}}, t_{\mathrm{f}}\rangle, F\rangle$, where t_{s} is the starting time of the interval and t_{f} is the finishing time of the interval in the time domain T_{abs}. F is the feature or property of the interval bounded by t_{s} and t_{f}.

From the perspective of the absolute-time domain, the feature *voiced* of figure 2 can be an absolute-time event $\alpha\tau\epsilon_1 = \langle\langle 113, 432\rangle, \mathrm{voiced_{phonation}}\rangle$.

Relative-time events and absolute-time events are governed by systems of axioms, and the basic rule of inference is *modus ponens*. The complete axiom set is given in Carson-Berndsen (1998, 73ff); a subset for each event type is given below for illustration. There are seven axioms governing the temporal relations between relative-time events. Those governing overlap ($\circ$) and precedence ($\prec$) are

$$R1 : \rho\tau\epsilon_i \circ \rho\tau\epsilon_i,$$
$$R2 : \rho\tau\epsilon_i \circ \rho\tau\epsilon_j \rightarrow \rho\tau\epsilon_j \circ \rho\tau\epsilon_i,$$
$$R3 : \rho\tau\epsilon_i \prec \rho\tau\epsilon_j \rightarrow \neg\rho\tau\epsilon_j \prec \rho\tau\epsilon_i.$$

There are 14 axioms governing the relations of temporal inclusion, overlap and precedence between absolute-time events. Examples of an axiom governing overlap and an axiom governing precedence of two absolute-time events, $\alpha\tau\epsilon_1 = \langle\langle t_{s1}, t_{f1}\rangle, F_1\rangle$ and $\alpha\tau\epsilon_2 = \langle\langle t_{s2}, t_{f2}\rangle, F_2\rangle$, are given in *A3* and *A5*, respectively:

$$A3 : t_{s1} \leqslant t_{s2} \wedge t_{f1} \geqslant t_{s2} \rightarrow \alpha\tau\epsilon_1 \circ \alpha\tau\epsilon_2$$
$$A5 : t_{f1} < t_{s2} \rightarrow \alpha\tau\epsilon_1 \prec \alpha\tau\epsilon_2.$$

Applying these axioms in the time domain T_{abs} to the events in the multilinear representation of figure 2, for example, would allow the inference to be made that the event with the feature *voiced* overlaps ($\circ$) the event with the feature *plosive*, and precedes ($\prec$) the event with the feature *voiceless*. There is a close relationship between absolute-time events and relative-time events; the latter is an abstraction of the former, that is to say, it represents the same facts but in a different temporal domain. Although the relative-time domain and the absolute-time domain are conceptually separate, they can be conflated for the purposes of implementation.

4. Finite-state models and event logics in speech recognition

The finite-state model and the event logic together form the basis for a constraint-based approach to phonological parsing based on the temporal interpretation of phonological categories as events, and using a flexible notion of compositionality based on underspecified structures with autosegmental tiers of parallel phonetic and phonological events. The overall architecture of the *Time Map* model in the context of linguistic word recognition is depicted in figure 3.

The autonomous feature-extraction component and the word parser provide the interfaces to the model. Features are detected individually from the speech signal by an autonomous feature-extraction component, which serves as a front-end to the phonological parser. The feature-extraction component does not classify over phones or phonemes but rather treats each feature as autonomous.

The features detected by the feature-extraction component are output together with their temporal annotations or boundary points and are represented in the figure in terms of parallel tiers of autonomous features; for illustrative purposes, only three tiers are shown. Each tier in this linear representation has its own temporal pattern or melody, and, therefore, the segmentation function across tiers is not the same as would be assumed by segmental phonologies. Rather than performing a segmentation into connected chart nodes, the *Time Map* solution to the *lattice-to-chart* problem is based on a mapping from the temporal annotations or boundary points to temporal relations between the hypotheses using an event logic. Parsing is then carried out entirely using the relations rather than the temporal annotations. Gaps are interpreted in terms of immediate precedence relations and overlapping properties are interpreted in terms of overlap relations. This approach is suitable for incremental processing.

As was discussed in the section on finite-state models, the phonotactic automaton of the phonological parser provides top-down constraints on the interpretation of the multilinear representations, specifying which overlap and precedence relations are expected by the phonotactics. Each time a final state of the automaton is reached, a well-formed syllable structure has been found. Since the input to the phonological parser is, in general, assumed to be underspecified due to noise, the *Time Map* model

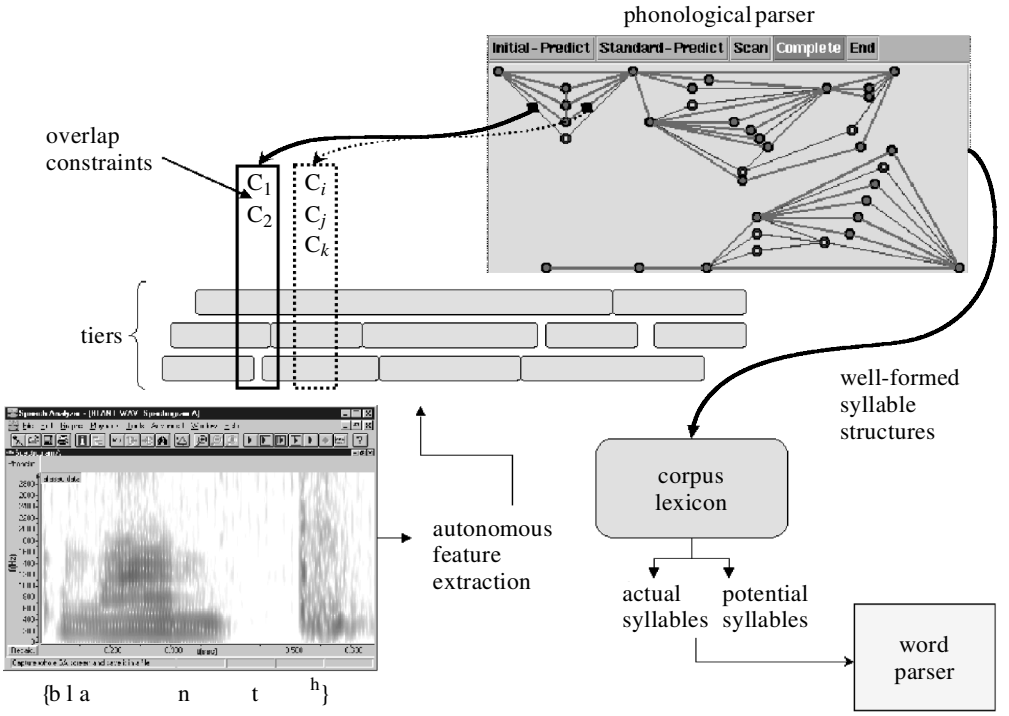

Figure 3. The *Time Map* model in the context of linguistic speech recognition.

must provide means of minimizing the discrepancy between the expectations defined in the top-down constraints and the actual data by allowing constraint relaxation and constraint enhancement. Constraint relaxation should be performed if only some of the constraints specified by the phonotactic automaton can be satisfied. Constraint enhancement should be performed to further specify the output if the constraints specify expectations that do not conflict with information found in the input. Clearly, constraint relaxation and constraint enhancement are interdependent and require a ranking of the constraints. This point will be discussed further in the next section. It is important to note, however, that the application of constraint enhancement does not guarantee that the output syllable structures are fully specified, only that they are well-formed.

Coarticulation is modelled by overlap of information in the multilinear representations. Since the overlap relations specified in the phonotactic constraints do not require features on different tiers to begin and end at the same time, a strict segmentation into non-overlapping units is not necessary. Phonological phenomena such as assimilations and elisions can, therefore, be represented in line with articulatory phonology in terms of feature (or gestural) overlap and magnitude. The degrees of overlap and magnitude are relevant for constraint relaxation and enhancement.

The phonological parser outputs well-formed syllable structures which are then input to a corpus lexicon. The corpus lexicon acts as a filter distinguishing between actual and potential syllables by assigning a higher confidence value to actual sylla-

bles of the corpus. These ranked syllable hypotheses are then passed, together with their temporal annotations, to the word parser. Since the syllable structures may be underspecified, the lexicon is feature based and, therefore, full specifications can be provided for corpus syllables if required. However, fully specified structures are not assumed by the word parser, since the main aim of this approach is to postpone a full specification in order to reduce the number of hypotheses.

The *Time Map* model has been implemented and tested within a linguistic word-recognition system, which is part of an experimental development environment for speech-recognition architectures. Within this system, the components can demonstrate differing interaction strategies and parameter settings for different types of analysis. Although performance was not the main concern of this work, a diagnostic evaluation has been performed with varying parametrizations of the system, which has led to some very promising results. As was mentioned above, the well-formed structures recognized at each level are in general underspecified, but full specifications in terms of phonemes or syllables can be calculated for evaluation. However, since this approach explicitly avoids a segmentation of the speech signal into non-overlapping units, a standard string-alignment evaluation procedure was not suitable for assessing the performance of the system.

A diagnostic evaluation procedure for the *Time Map* model was developed that consisted of a logical evaluation, with respect to a data model, and an empirical evaluation, with respect to real signal data. The logical evaluation was responsible for testing the soundness and completeness of the parser and knowledge components of the model with fully specified data, and the empirical evaluation tested the performance of the model in the context of the complete word-recognition system. In addition to a string-alignment evaluation, a *Time Map* evaluation procedure was performed. With one particular parametrization of the system, a phoneme recognition rate of 66.97% and a syllable recognition rate of 35.19% were attained on many-speaker utterances of spontaneous scheduling task dialogues. Diagnostic evaluation showed that the relatively low syllable-recognition rate was due to one or two features being unreliable, leading to the complete syllable not being recognized. Furthermore, since phoneme recognition was not a task of the phonological parser, the phoneme rate was calculated on the basis of the recognized syllables. However, the results are remarkable for a purely knowledge-based system and it is anticipated that improvements will be made by more efficient constraint-relaxation and constraint-enhancement procedures, as will be discussed in the next section. Further details of the evaluation procedure and the results can be found in Carson-Berndsen (1998).

5. Integrating statistics

The *Time Map* model is a formally specified linguistic–symbolic approach with fine granularity which offers a solution to the problem of projecting onto potentially well-formed structures at the phonetics–phonology interface in speech recognition. However, in the same way that the fine-grained knowledge of this model has been used to fine-tune stochastic models with considerable success, so it is also possible to use statistical information to fine-tune the *Time Map* approach for specific applications. The rest of this paper identifies the areas in which statistics can play a role and discusses these with respect to the overall architecture. There are three main integration areas for statistics in the *Time Map* model which differ in granularity.

The first integration area is constraint ranking, which represents the lowest level of granularity in that the constraints refer to individual temporal relations. The second area of integration is in connection with the weighting of the phonotactic automaton. Automaton 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.

The notion of constraint ranking was seen in the previous section to play an important role in constraint relaxation and constraint enhancement. Constraint ranking for this model 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. The rest of this section will only be concerned with such data-oriented ranking, although, clearly, it is more likely that a combination of the factors will be appropriate for constraint ranking, and, therefore, a free parametrization of the system should allow adequate parameters to be chosen that define a compromise between maximal recognition rates and minimal analysis overhead.

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 standard deviation, for example. Constraint enhancement can also be applied according to such a ranking when a frequent feature is expected but not present in the input but there is nothing in the input that would exclude it. This method of constraint ranking for relaxation and enhancement has yet to be integrated into the model, but this is the integration area for statistics which could lead to reductions in the error rate. The other two integration areas for statistics will influence the size of the search space and the confidence of the hypotheses and, thus, improve accuracy by reducing the number of false positives.

The second area of integration of statistics in the *Time Map* model is in connection with corpus-based fine-tuning of the phonotactic automaton. Using corpus-based statistics, weights can be calculated for the individual transitions in the automaton, and, in line with proposals made by Pereira & Riley (1996), the total weight can be defined, using commutative semirings as a formal basis, as the extension of the weights of the corresponding paths. In addition to determinization and minimization, which make the automaton more efficient from a processing viewpoint, it may also be interesting to experiment with specific data-oriented phonotactic models which are generated with respect to particular corpora. For example, Belz (1998) presents an approach for the automatic acquisition of finite-state models of phonotactic constraints based on genetic algorithms. The composition of the data-oriented automaton and the *Time Map* automaton would be another method of assigning a corpus-based weighting to the phonotactic automaton. This is currently being investigated.

The third area of integration for statistics is the lexicon, and this is the component of the *Time Map* model that has made the most use of statistical information thus far. Since the lexicon specifies only those syllables that are in the corpus, it has been possible to include in the syllable entries statistical information such as frequency, average duration and standard deviation, both of the syllable as a whole and of the duration based on the average durations of its individual parts taking context infor-

mation into account. Currently, this is the only place in the *Time Map* model where a corpus-based ranking is performed by providing actual (corpus) syllables with a higher confidence value than potential syllables, and a more fine-grained ranking of hypotheses within the category of actual syllables is performed by allowing syllable hypotheses with the greatest degree of specification and which most closely match the average durations to be ranked higher than their counterparts. This approach is currently under development in the context of an ongoing research project.

This section has been concerned with the integration of statistics into the *Time Map* model. It has been shown with respect to the overall architecture of the model that there are a number of areas in which statistics can play a role. In all that has been said here, however, it should be obvious that statistical information is regarded not as a basis for converting the linguistic–symbolic model to a stochastic one but as an aid for application-specific tuning.

6. Conclusion

This paper has been concerned with a computational linguistic model of phonological interpretation which provides a framework in which multilinear parallel event representations of speech utterances are temporally interpreted using finite-state machines. Rather than concentrate purely on system performance and recognition results, the aim of this approach has been to develop a principled, formally specified linguistic theory of phonological interpretation which investigates the role played by symbolic constraints on well-formedness and provides important fine-grained knowledge representations for speech-technology applications. The informational structures used by this system could on the one hand be enhanced by statistical information for application-specific tuning but also be made available to other stochastic word-recognition systems for the purposes of structural fine-tuning. This approach is directly relevant to multi-sensor input applications (see Carson-Berndsen 1999*a*, *b*), since the finite-state methodology and the event logic provide an overall framework for the temporal interpretation of parallel structures.

The *Time Map* model addresses a number of important issues that have arisen in connection with phonological theories and speech recognition. Firstly, it provides a computational linguistic solution to the new word problem in speech recognition by using a complete finite-state phonotactics of the language to define the notion of well-formedness. Secondly, it provides the phonological description with a temporal interpretation in terms of an event logic which not only deals with abstract phonological examples but also with concrete speech tokens. Thirdly, coarticulation and other phonological processes are modelled in a multilinear representation by providing an interpretation of overlap and gap phonomenon avoiding a rigid segmentation into non-overlapping phonemes. Fourthly, the model copes with underspecification by employing constraint relaxation and constraint enhancement. Finally, since the model is knowledge based, it has not been pre-tuned to any particular speaker or any particular corpus, although ways in which this could be done using statistical information were discussed.

The model has been fully implemented and tested in an incremental linguistic speech-recognition system. In that system, the *Time Map* model provided the formal foundation for the phonological parsing component, which was interfaced with an autonomous feature-extraction component (Hübener & Carson-Berndsen 1994) and

a word parser. However, since the model is independent of these other components, it should be immediately possible to interface it directly—or use it in parallel—with other multilinear recognition systems, such as those suggested by Kirchhoff (1996) or King *et al*. (1998), in which the paradigms being followed are the hidden Markov approach and the connectionist approach, respectively. Although the *Time Map* model was developed originally in the phonetic and phonological knowledge domains, it does generalize to higher levels in the prosodic hierarchy, thus avoiding the need for a level-specific segmentation.

References

Allen, J. F. 1983 Maintaining knowledge about temporal intervals. *Comm. ACM* **26**, 832–843.

Belz, A. 1998 An approach to the automatic acquisition of phonotactic constraints. In *Proc. SIGPHON 98: The Computation of Phonological Constraints* (ed. T. M. Ellison), pp. 34–55.

Bird, S. 1995 *Computational phonology: a constraint-based approach*. Cambridge University Press.

Bird, S. & Ellison, T. M. 1994 One-level phonology: autosegmental representations and rules as finite state automata. *Comp. Ling.* **20**, 55–90.

Bird, S. & Klein, E. 1990 Phonological events. *J. Ling.* **26**, 33–56.

Carson-Berndsen, J. 1998 *Time map phonology: finite state models and events logics in speech recognition*. Dordrecht: Kluwer.

Carson-Berndsen, J. 1999*a* A generic lexicon tool for word model definition in multimodal applications. *Proc. EUROSPEECH 99, Budapest, Hungary*, vol. 5, pp. 2235–2238.

Carson-Berndsen, J. 1999*b* A feature geometry based lexicon model for speech applications. In *Proc. IDS 99*, pp. 65–68.

Freksa, C. 1992 Temporal reasoning based on semi-intervals. *Artificial Intelligence* **54**, 199–227.

Gibbon, D. 1992 Prosody, time types and linguistic design factors in spoken language system architectures. In *KONVENS 92, 1. Konferenz 'Verarbeitung natürlicher Sprache'* (ed. G. Görz), pp. 90–99. Springer.

Hübener, K. & Carson-Berndsen, J. 1994 Phoneme recognition using acoustic events. In *Proc. ICSLP 94, Yokohama, Japan*, vol. 4, pp. 1919–1922.

Jusek, A., Rautenstrauch, H., Fink, G. A., Kummert, F., Sagerer, G., Carson-Berndsen, J. & Gibbon, D. 1994 Dektektion unbekannter Wörter mit Hilfe phonotaktischer Modelle. In *Mustererkennung 94, 16. DAGM-Symposium Wien*, pp. 238–245. Springer.

Kaplan, R. M. & Kay, M. 1994 Regular models of phonological rule systems. *Comp. Ling.* **20**, 331–378.

Kay, M. 1987 Nonconcatenative finite-state morphology. In *Proc. EACL 87, Copenhagen*, pp. 2–10.

King, S., Stephenson, T., Isard, S., Taylor, P. A. & Strachan, A. 1998 Speech recognition via phonetically featured syllables. In *Proc. ICSLP 98*, pp. 1031–1034.

Kirchhoff, K. 1996 Phonologisch strukturierte HMMs. In *Natural language processing and speech technology* (ed. D. Gibbon), pp. 55–63. Berlin: Mouton de Gruyter.

Kornai, A. 1995 *Formal phonology*. Levittown, PA: Garland.

Koskenniemi, K. 1983 *Two-level morphology: a general computational model for word-form recognition and production*. University of Helsinki, Department of General Linguistics Publications, no. 11.

Pereira, F. C. N. & Riley, M. D. 1996. Speech recognition by composition of weighted finite automata. CMP-LG archive paper 9603001.

Smolensky, P. 1999 Who's afraid of linguistic theory? In *elsnews, June 1999*, pp. 6–7.

Van Benthem, J. F. A. K. 1983 *The logic of time*. Dordrecht: Reidel.

Wagner, P. 1997 Phonologie und automatische Sprachsynthese. MA thesis, University of Bielefeld.

Discussion

S. J. Young (*University of Cambridge, UK*). The main focus of your presentation was on the recognition of new words. But are new words hard to recognize? It seems to me that the most challenging outstanding problems in speech recognition lie not in word models but in language modelling. An important useful property of the standard statistical speech-recognition architecture is *bootstrapping*, that is, using a model trained on a small dataset to annotate a much larger dataset, which can then be used for training a larger model.

J. Carson-Berndsen. Using phonotactic models reduces the need for training our speech recognizer on a large corpus, because of the linguistic generality of the constraints. For example, we noticed that a consonant sequence defined by our German model, /-lnst/, did not occur in any word in the dictionary. But intuition suggested it would be phonotactically possible in, for example, the coinage *kölnst du?*, *Do you cologne?*, which native speakers, when asked, accepted as well-formed.

D. Kazakov (*University of York, UK*). How well does this approach perform compared with methods based on analogy? For example, one accepts *blant* as a likely word, because one knows that *plant, plaster* and *bluster* are words.

J. Carson-Berndsen. The two approaches are complementary: the features and multilinear representations of my system could be employed in the computation of analogy.

H. Alshawi (*AT & T Laboratories, Florham Park, NJ, USA*). For the recognition of foreign names and loan-words, do you envisage a universal phonotactic model, or several language-specific models applied in parallel?

J. Carson-Berndsen. Since we are already using different models for different languages, we could use several of them as parallel sub-models in that case. I do not think a single universal model would be feasible.

F. Pereira (*AT & T Laboratories, Florham Park, NJ, USA*). Another advantage of the standard speech-recognition architecture is that multiple sources of uncertainty can be resolved by exploiting a mathematically well-defined combination of probabilistic models of different levels of structure. For example, voicing detection can be rather unreliable, but the word model could ameliorate the uncertainty of the acoustic analysis.

J. Carson-Berndsen. Our model has three mechanisms for coping with such uncertainty. First, there is the differential weighting of constraints and the relaxation of, for example, infrequency constraints. Second, by using underspecified representations, variable or uncertain features can be omitted. Third, we try to defer decisions until late in the analysis process, to avoid incorrect commitment at an earlier stage. In this way, for example, phonotactic constraints could fill in information missing from the signal analysis.