# EVENT RELATIONS AT THE PHONETICS/PHONOLOGY INTERFACE JULIE CARSON-BERNDSEN DAFYDD GIBBON Universität Bielefeld Fakultät für Linguistik und Literaturwissenschaft Postfach 8640 4800 Bielefeld 1 Germany

# Summary

In this paper a procedure for the construction of event relations at the phonetics/phonology interface is presented. The approach goes further than previous formal interpretations of autosegmental phonology in that phonological relations are explicitly related to intervals in actual speech signals as required by a speech recognition system. An event structure containing the temporal relations of overlap, precedence and inclusion is automatically constructed on the basis of an event lattice with time annotations derived from the speech signal. The event structure can be interpreted linguistically as an autosegmental representation with assimilation, long components or coarticulation. The theoretical interest of this work lies in its contribution to the solution of the projection problem in speech recognition, since a rigid mapping to segments is not required.

#### 1. Motivation

In the processing of speech one of the major problems is the projection problem at the phonetics/phonology interface: sounds and words are realised with different degrees of coarticulation (overlap of properties) in different lexical, syntactic, and phonostylistic contexts and thus a segmentation into phonemes alone is too rigid in order to capture all variants. Furthermore, the set of possible words in natural languages, analogous to the set of sentences, is infinite. In fact, even subsets of these sets may be so large that a simple list is no longer tractable. This has so far proved to be an insuperable problem for the simple concatenative word models of current speech recognition systems, whether phoneme, disyllable, or word based. In this paper, a new approach to this problem is proposed, starting from recent wellmotivated developments in phonology such as autosegmental phonology (Goldsmith, 1976,1990), articulatory phonology (Browman & Goldstein, 1986,1989), underspecification theory (Archangeli, 1988; Keating, 1988) and phonological events (Bird & Klein, 1990). The overall context for the work presented here is a further development of the PhoPa system (Carson, 1988; Carson-Berndsen, 1990) for phonological word parsing with a feature-based phonotactic net. The present approach goes beyond these studies in deriving

phonological relations directly from speech data, and in providing detailed language-specific top-down phonotactic constraints.

For phonological parsing a flexible notion of compositionality is utilised based on underspecified structures with 'autosegmental' tiers of parallel phonological events which avoid a rigid mapping from phonetic parameters to simple sequences of segments. The motivation for using an event-based phonological representation was to use phonological knowledge as represented in the phonotactic net (thus also maintaining the notion of underspecification and optimisation by the use of feature cooccurrence restrictions) while catering for those phenomena arising in continuous speech which do not correspond to the phonotactics of the language. An example of this kind of phenomenon found during the labelling of the EUROM-0 speech data in the SAM project (ESPRIT 2589 cf. Braun, 1991b) is the cluster [szs] in the German word [vE:RUNszste:m] as a pronunciation of /vE;RUNszYste:m/ Währungssystem (see section 3).

By using a phonotactic description based on an autosegmental representation of events and the temporal relations which exist between them, a rigid segmentation at the phonetic level is no longer necessary. A further advantage of an event representation with temporal annotations at the phoneticsphonology interface concerns the exchange of differing types of information between the two levels. An event is interpreted as an interval with a particular property, and it is not necessary to confine the possible set of properties to conventional phonological features such as voice or nasal but acoustic properties of actual speech signals such as "frication noise" or "syllable peak" may be included.

### 2. Event Relations

Three stages are involved in the determination of signal-derived event relations at the phonetics/phonology interface. These are: (1) Event Detection, which will be discussed from the point of view of phonetic and phonological levels of representation in section 2.1., (2) Event Mapping where the relations between the individual events are constructed automatically, which is discussed in section 2.2 and (3) Event Structure Constraints, defining phonological structure, which are discussed in section 2.3. The work described here is concerned primarily with speech recognition rather than synthesis and in particular with its phonological parsing component as opposed to the acoustic front end. The event relations generated at the phonetics/phonology interface serve as input to a constraint-based phonological parser whose knowledge base is an event-based description of the phonotactics of the language.

#### 2.1. Phonetic and Phonological Events

Assuming that the feature detectors at the acoustic level recognise events each consisting of a property and an interval together with a measure of confidence, it is possible to define a procedure which automatically constructs temporal relations of overlap, precedence and inclusion over intervals Bird & Klein (1990) have some reservations about the use of endpoints of intervals at the phonological level. However, absolute temporal annotations must indeed be provided at the phonetic level on the basis of threshold and confidence values for a particular acoustic event in a speech signal token, and the use of these in the calculation of temporal relations for a given signal within an actual speech recognition procedure is in fact necessary, not an option.

At the phonological level, an event is simply a pair of a property and an interval <P, I>. At the phonetic level, an event is a quadruple <P, ts, tF, C>, providing information on event-type (property), start of interval, end of interval and confidence value. This serves as input to event mapping. The output of the mapping is a set of tuples  $\langle e_i, R, e_i \rangle$  where  $e_i$  and  $e_i$  represent events and R is the temporal relation which exists between them (overlap, precedence or temporal inclusion). Using phonological constraints based on simplex and complex phonolological event structures, the phonologically relevant information is abstracted from this set of tuples. It is not the temporally annotated events themselves which are interesting for the phonological parser but the temporal relations which exist between these events (cf. section 2.3).

### 2.2. Event Mapping

In the speech recognition context there is a mapping of absolute phonetic events to abstract temporal relations between events is described. The algorithm for the automatic construction of event relations has the following properties: Each event pair is tested only once; there is no explicit statement of reflexivity. The reflexivity and symmetry of overlap are not reflected in the output, but can be inferred by Modus Ponens from the axioms at the phonological level. Inclusion is a special case of overlap; thus, when an event is temporally included in another these events also overlap, and the algorithm makes use of this fact. There are nine types of overlap, seven of which are instances of inclusion, and all are catered for by the algorithm. It was found that the relation of temporal inclusion played an important role in the constraints needed for phonological parsing (Carson-Berndsen, 1991). Simultaneity was not considered due to the fact that phonetic decisions are made on the basis of confidence values and thus the likelihood of true simultaneity is low. There is no difficulty, however, in augmenting the algorithm to cater for this if required since it is in fact a relationship of mutual temporal inclusion.

The relations of overlap and precedence which hold between pairs of events are governed by a set of axioms; event structures are defined as a collection of events and constraints. These axioms can be regarded as having three different functions: inference, abbreviation and consistency checking.

With respect to the abbreviation function of the axioms, this feature is not currently availed of in the algorithm as this would not reduce the search space. The consistency checking function of the axioms would be an extra step after the relations have been constructed. The output of the event mapping is an event lattice, analogous to the traditional disjunctive lattices of phoneme, syllable or word-based speech recognition, but not so far considered in previous work based on autosegmental structures.

#### 2.3. Event Structure Constraints

There is clearly no direct correspondence between events as measured in a signal, and abstract phonological structures. These levels differ in five major ways: first, the signal-derived relations may be incomplete, owing to noisy input; second, the signalderived event relations approximate to the transitive closure of the phonologically relevant minimal specification of the event structure, and must therefore be reduced by appropriate criteria; third, contextually conditioned phonetic reductions, assimilations and epentheses must be resolved; fourth, explicit complex phonological structures need to be defined; fifth, there may be no simple relation between event endpoints and nodes in parse chart structures. To complete the mapping from phonetic events to phonological event structures, constraints must be formulated which fulfil these tasks. The third type will be briefly discussed in section 3; the rest of the present section is mainly concerned with the fourth type. For the phonological component in the present system, a distinction is made between simplex and complex events.

A simplex phonological event is defined as the basic unit of input from the phonetic component; at the phonetic level these events are in general a function of several parameters and are therefore by no means 'simplex' at this level. A complex phonological event is constructed compositionally in terms of the precedence, overlap and inclusion relations at the phonological level. So for example the composition of the simplex events occlusion, transient and noise results in the complex event plosive. Complex events also refer to larger structures relevant at the phonological level such as svilable onset or reduced svilable. Using the constraint axiom set, further relations between these complex events are inferred. In the speech recognition context, absolute speech signal constants are required to be assigned to the largest complex events in order to permit synchronisation at higher levels. The output of constraint application is thus a complex event lattice which is subsequently mapped to a linguistic parse chart (cf. Chien & al. 1990).

# 3. An Example

In this section, an example of input and output in the system for generating the relations between phonetic events in a token of the English word palm /pa:m/ is discussed (cf. also Carson-Berndsen, 1991). The speech signal is shown in Figure 1; the phonemic annotations and display were produced with the SAMLAB speech signal labelling system (Braun, 1991a). The events used in this analysis are based on a feature set proposed by Fant (1973); although the features have labels which indicate articulatory features, they are in fact acoustically based. A diagrammatic representation of the detected events in an approximately 520 msec interval is shown in Figure 2. The temporally annotated events are passed to the phonological component of the speech recognition system in the interface format given in (3). Before the above algorithm is applied, the tuples are uniquely identified and translated into a variety of attribute-value notation as shown in (4) (note that confidence values are not considered further here).

- (3) Temporal input from the phonetic level <voiceless, 0, 91.19, C> <voiced, 91.2, 517.5, C> <glide, 452.6, 498.2, C> <occlusive, 0, 35.4, C> <transient, 34.5, 60.6, C> <noise, 60.61, 91.16, C> <vowellike, 94.29, 392.6, C> <nasal, 402.9, 518.6, C> <bilabial, 20.45, 93.2, C> <tongue-retracted, 93.21, 392.6, C> <bilabial, 392.62, 518.2, C>
- (4) Event inventory
  - VOI(voiceless, <0.91.19>)e<sub>i</sub>:
  - VOI(voiced, <91.2,517.5>)  $e_2$ :
  - GLI(glide, <452.6,498.2>)  $e_3$ :
  - OCC(occlusive, <0.35.4>)e₄:
  - TRA(transient, <34.5, 60.6>)с.:
  - NOI(noise, <60.61,91.16>) e6:
  - VOW(vowellike, <94.29,392.6>) e7:
  - NAS(nasal, <402.9,518.6>) C8:
  - LAB(bilabial, <20.45,93.2>) co:
  - e<sub>10</sub>: TON(retracted, <93.21,392.6>)
  - e<sub>11</sub>: LAB(bilabial, <392.62,518.2>)

Of particular interest to the phonological parser are the precedence relations between those event properties of the same type and the overlap and temporal inclusion relations between event properties of differing types. Initially all relations between the individual events are generated automatically in (5). The temporal relations of overlap, precedence and inclusion are represented by the symbols "", '<' and '{' respectively.

One of the motivations for having chosen an event-based phonology for coping with the interface between phonetics and phonology was to be able to cater for phenomena which do not correspond to the phonotactics of the language. It may be the case, as given in the example Währungssystem in section 2, that the information on the centre portion of the signal, which is shown in (6) after the translation into attribute-value structure, is provided by the phonetic component.

(6) Temporal annotations for [szst] cluster e1: FRICATION(fricative, <0, 301.3>)  $e_{1}$ : VOICE(voiced, <79.9, 229.3>) e<sub>3</sub>: VOWELLIKE(vowellike, <128.5, 202.6>) e.: OCCLUSION(occlusive, <301.31, 334.6>)

There is not a full match between the output of the event mapping and any phonological representation, because FRICATION is continuous throughout and and thus overlaps VOWELLIKE rather than both preceding and following it. However, the phonological constraints include information on possible phonotactic structures; these will not be discussed here in detail (but cf. Carson-Berndsen 1992). Positions in these structures are underspecified in terms of events, thus indirectly defining a priority between specified and nonspecified event types at those positions. In this case, at the relevant VOWELLIKE interval FRICATION overlap is not specified, and thus a phonotactic match is permitted; VOICE is also not specified for initial sibilants. Note that vowel quality does not need to be specified in detail in the phonotactics. If an actual lexical item is more highly specified at these positions. it will match this part of the phonotactic structure, thus ultimately allowing the relevant portion of phonological representation of Währungssystem to be derived.

- (7) Constraints for [szst] cluster (fragment)
- (explicitly required by phonotactics)  $\mathbf{e}_1 < \mathbf{e}_4$
- $c_2 < c_4$ (explicitly required by phonotactics)
- $c_3 < c_4$ (explicitly required by phonotactics)
- (not specified by phonotactics)
- $e_1^{\circ} e_2^{\circ}$  $e_1^{\circ} e_3^{\circ}$ (not specified by phonotactics)
- c2° c3 (explicitly required by phonotactics)



Figure. 1

voiceless				
voiced				
occlusion	<u> </u>			
transient				
noise				
vowellike				
nasal				
bilabial			······································	
tongue retract	.ed			
	p	a:	n m	
		Figure 2		
) Output of Automa	tic Event Mapping	-		
<e1, <,="" e2=""></e1,>	<e1, <,="" e3=""></e1,>	<e1, e4="" °,=""></e1,>	<e1, e4="" {,=""></e1,>	
<e1, e5="" °,=""></e1,>	<e1, e5="" {,=""></e1,>	<e1, e6="" °,=""></e1,>	<e1, e6="" {,=""></e1,>	
<e1, <,="" e7=""></e1,>	<e1, <,="" e8=""></e1,>	<c1, e9="" °,=""></c1,>	<e1, <,="" e10=""></e1,>	
<e1, <,="" e11=""></e1,>	<e2, e3="" °,=""></e2,>	<e2, e3="" {,=""></e2,>	<e4, <,="" e2=""></e4,>	
<e5, <,="" e2=""></e5,>	<e6. <.="" e2=""></e6.>	<e2. e7="" °.=""></e2.>	<e2. e7="" {.=""></e2.>	
<e2. e8="" °.=""></e2.>	<e2. e9="" °.=""></e2.>	<e2. e10="" °.=""></e2.>	<e2 e10="" {=""></e2>	

, ,			, 02, [, 010,
<e2, e11="" °,=""></e2,>	<e4, <,="" e3=""></e4,>	<e5, <,="" e3=""></e5,>	<e6, <,="" e3=""></e6,>
<e7, <,="" e3=""></e7,>	<e8, e3="" °,=""></e8,>	<e8, e3="" {,=""></e8,>	<e9, <,="" e3=""></e9,>
<e10, <,="" e3=""></e10,>	<e11, e3="" °,=""></e11,>	<e11, e3="" {,=""></e11,>	<e4, e5="" °,=""></e4,>
<e4, <,="" e6=""></e4,>	<e4, <,="" e7=""></e4,>	<e4, <,="" e8=""></e4,>	<e4, e9="" °,=""></e4,>
<e4, <,="" e10=""></e4,>	<e4, <,="" e11=""></e4,>	<e5, <,="" e6=""></e5,>	<e5, <,="" e7=""></e5,>
<e5, <,="" e8=""></e5,>	<e9, e5="" °,=""></e9,>	<e9, e5="" {,=""></e9,>	<c5, <,="" e10=""></c5,>
<e5, <,="" e11=""></e5,>	<e6, <,="" e7=""></e6,>	< c6, <, c8>	<e9, e6="" °,=""></e9,>
<e9, e6="" {,=""></e9,>	<c6, <,="" e10=""></c6,>	<e6, <,="" c11=""></e6,>	<e7. <.="" e8=""></e7.>
<e9, <,="" e7=""></e9,>	<e10, e7="" °,=""></e10,>	<e10, e7="" {,=""></e10,>	<e7. <,="" e11=""></e7.>
<e9, <,="" e8=""></e9,>	<e10, <,="" e8=""></e10,>	<c8. c11="" °.=""></c8.>	< e9. <. e10 >
<e9, <,="" c11=""></e9,>	<e10. <.="" e11=""></e10.>		, ,
, ,			

# 4. Final Remarks

In this paper a new solution to the projection problem in speech recognition is proposed in the form of a three-stage procedure for the automatic construction of event relations and phonological event structures, starting with an event lattice of simplex events in the form of temporal annotations provided by the acoustic phonetic component of a speech recognition system. In contrast to the purely concatenative solutions to word compositionality which are conventionally used, the present flexible approach using the three compositional relations of overlap, precedence and temporal inclusion promise a principled and effective solution to the projection problem at the phonetics/phonology interface.

# 5. Bibliography

Bird, S; E. Klein (1990): Phonological Events, In: Journal of Linguistics 26, 33-56.

Braun, G. (1991a): SAMLAB. Ms. University of Bielefeld.

Braun, G. (1991b):

Tools in Speech Technology: Problems in Segmental Labelling. Paper held at the Workshop on Computational (Morpho)Phonology, ZiF, Universität Bielefeld, 23-25 October 1991.

Browman C.P.; L. Goldstein (1986): Towards an articulatory phonology. In: Phonology Yearbook 3:219-252

Browman C.P.; L. Goldstein (1989):

Articulatory gestures as phonological units. In: *Phonology 6*, Cambridge: Cambridge University Press, 201-251

Carson, J. (1988):

Unification and Transduction in Computational Phonology. In: Proceedings of the 12th International Conference on Computational Linguistics, Budapest, 106-111.

Carson-Berndsen, J.; D. Gibbon; K. Knäpel (1989): Interim Report 31.03.89 and Final Report 30.09.89 Forschungsprojekt: Entwicklung phonologischer Regelsysteme und Untersuchungen zur Automatisierung der Regelerstellung für Zwecke der automatischen Spracherkennung. Forschungsprojekt finanziert von der Deutschen Bundespost, Ms. Universität Bielefeld Carson-Berndsen, J. (1990):

Phonological Processing of Speech Variants. In: Proceedings of the 13th International Conference on Computational Linguistics, Helsinki 3:21-24 Carson-Berndsen, J. (1991):

Ereignisstrukturen für phonologisches Parsen. Project Report: ASL-TR-9-91/UBI, University of Bielefeld, August 1991

Carson-Berndsen(1992):

An event-based phonotactics for German. ASL-TR-29-92/UBI, University of Bielefeld, February 1992

Chien, L-F., K.J.Chen, L-S.Lee (1990) .:

An augmented chart data structure with efficient word lattice parsing scheme in speech recognition applications. *COLING 90*, Vol. 2: 60-65. Helsinki.

Fant, G. (1973):

Speech Sounds and Features. Cambridge, Massachusetts: MIT Press.

Goldsmith J.(1976):

Autosegmental Phonology. Bloomington, Indiana: Indiana University Linguistics Club.

Goldsmith, J. (1990):

Autosegmental and Metrical Phonology. Cambridge, Massachusetts: Basil Blackwell Inc.

Note: The work presented in this paper was financed by the German Ministry for Research and Technology within the project Architectures for Speech and Language Systems (VERBMOBIL-ASL-Nord).