BMBF

Verbmobil
Verbundvorhaben

Parametrised Phonological Event Parsing

Julie Carson–Berndsen,
Guido Drexel

Universität Bielefeld

Vm

Report 172
September 1996

September 1996

Julie Carson–Berndsen, Guido Drexel
Universität Bielefeld (UBI)
Fakultät für Linguistik und Literaturwissenschaft
Universitätsstr. 25
Postfach 10 01 31
33501 Bielefeld
Tel.: (0521) 106 - 3510
Fax: (0521) 106 - 6008
e-mail: {berndsen,drexel}@Spectrum.Uni-Bielefeld.DE

Gehört zum Antragsabschnitt: 15.6 Interaktive Phonologische Interpretation

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 B 2 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Abstract

This paper describes a phonological event parser for spoken language recognition which has been provided with a parametrisable development environment for examining the extent to which linguistically significant issues such as linguistic competence (structural constraints) and linguistic performance (robustness) can play a role in the spoken language recognition task.

Ein phonologischer Ereignisparser zur Erkennung gesprochener Sprache wird zusammen mit einer parametrisierbaren Entwicklungsumgebung vorgestellt. Diese Umgebung dient nicht nur der Entwicklung und Konsistenz– und Vollständigkeitsprüfung des zugrundeliegenden computerphonologischen Modells, sondern ermöglicht auch eine gezielte Evaluierung ausgewählter linguistisch motivierter *constraints* zur robusten Erkennung gesprochener Sprache.

1 Introduction

This paper[1] describes a knowledge–based phonological event parser for spoken language recognition which has been provided with a parametrisable development environment for examining the extent to which linguistically significant issues such as linguistic competence (structural constraints) and linguistic performance (robustness) can play a role in the spoken language recognition task. The primary aim of the work presented here is to go beyond a holistic performance evaluation of components of a spoken language recognition system and enable a diagnostic evaluation of independent parameters. This method of diagnostic evaluation can provide more insights into the role played by linguistically motivated constraints in spoken language recognition than a simple indication of minor changes in recognition rate provided by the holistic approach. Linguistic competence of the native speaker concerns the construction of knowledge bases of linguistic constraints which allow a recogniser for spoken language to distinguish between *actual* structures (i.e. in the lexicon) and *potential* structures (i.e. new words) and to reject ill–formed structures based on information other than the fact that they are not in the lexicon. This is not to claim, however, that a spoken language recognition system should be purely knowledge–based. The long term aim of the work presented here is to provide a linguistic basis for combining knowledge–based and the stochastic approaches to spoken language recognition. In fact, it has already been shown that the explicit incorporation of phonological knowledge can provide useful structural constraints for the fine tuning of stochastic models (Jusek et al., 1994). In connection with the model presented in the next section the primary aim has been to examine the role of purely knowledge–based constraints, postponing the incorporation of stochastic constraints until a complete diagnostic evaluation of isolated knowledge parameters has been undertaken. The development of a model of native speaker linguistic performance clearly presupposes a model of native speaker linguistic competence, since it assumes the *a priori* definition of a norm, from which the native speaker may be allowed to deviate during normal speech. In fact a closer look at spoken language shows that the norm is seldom adhered to; fast speech is a series of words, hesitations, pauses, incomplete utterances, corrections, all of which can be filtered by the listener and mapped via a model of native speaker linguistic competence onto an interpretable representation. Quite a lot of research has been done in the area of native speaker performance ranging from work on syntactic normalisation (cf. for example Langer, 1990) to a computational treatment of rule–based speech

[1] This paper was originally published in Dafydd Gibbon (ed.): *Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference. Bielefeld, October 1996*, pp. 64–70. Berlin, etc.: Mouton de Gruyter.

variation (cf. for example Carson-Berndsen, 1990; Kirchhoff, 1995). It is the issue of native speaker performance which is closely related to the notion of robustness of a system.

2 Phonological event parsing

The basic computational model for phonological event parsing developed within the context of research into the architectures of speech and language systems in the Verbmobil research project operationalises a novel, constraint–based approach to phonological parsing based on temporal interpretation of phonological categories as events. It is one of three components which participates in the linguistically based word recognition system BELLEx3 (cf. Althoff et al., 1994, 1995; Hübener and Carson-Berndsen, 1994). The phonological event parser uses a flexible notion of compositionality in line with recent developments in multilinear phonology (cf. Goldsmith, 1976; Bird and Klein, 1990; Carson-Berndsen, 1993) based on underspecified structures with 'autosegmental' tiers of parallel phonological events which avoids a rigid mapping from acoustic parameters to simple sequences of phoneme segments. Independent acoustic events contribute different information relevant for the composition of phonological events. The following is a brief description of the phonological event parser:

1. Goal: to construct phonological and syllable event hypotheses from an acoustic event lattice and, in doing so, to restrict the search space of other components in the spoken language recognition system.
2. Input: a lattice of acoustic events classifying features of manner and place of articulation, phonation and vowel quality which are detected by an acoustic event recogniser (cf. Hübener, 1993; Hübener and Carson-Berndsen, 1994) on the basis of the speech signal.
3. Output: a lattice of phonotactically well–formed syllable and sub–syllable events which are passed to a word parser for structured word recognition.
4. Knowledge Base: a phonotactic constraint network which is defined in autosegmental representation with respect to a primary timing tier (cf. Carson-Berndsen, 1992, 1993). Each element of the primary tier defines constraints on overlap and immediate precedence of phonological events in a particular syllable position, e.g. vowellike overlaps front, vowellike precedes fricative. The primary autosegmental tier is represented as a finite–state automaton which interprets these constraints in line with current work in finite–state phonology (cf. Kaplan and Kay, 1994).

5. Basic Method: the phonological event parser employs a basic iterative algorithm structure of PREDICT, SCAN and COMPLETE. The general search strategy is breadth–first with an *everywhere–predict*.[2] SCAN is based on the interpretation of constraints on temporal relations as defined by the primary tier of the phonotactic knowledge base which provides control and top–down constraints for the input acoustic event representation. COMPLETE is undertaken when a well–formed and, in the general case, underspecified representation of the syllable structure has been found in accordance with the constraints.

3 Parametrisation and functionality

The event parser can be parametrised according to the following basic classification: knowledge base parameters, performance parameters, predict strategy parameters, temporal parameters, communication parameters and system parameters. It would be possible also to include a parameter responsible for regulating the extent of stochastic information used by the parser. However, although this parameter has been included at the conceptual level of the development environment, it has not yet been operationalised. All parameters affect the performance of the system either in terms of speed or in terms of recognition rate and each of the parameter types contribute to different aspects of the functionality of the system as a whole. The *knowledge base parameters* define the application of constraints on linguistic competence of the model which, in contrast to the other parameter types, can be defined independently of the processing system on the basis of strictly linguistic empirical research. Such constraints must offer a solution to the projection problem at the phonetics/phonology interface (cf. Carson-Berndsen and Gibbon, 1992) which concerns the variability and the compositionality of speech. Variability refers to the fact that sounds and words are realised with different degrees of coarticulation in different contexts. Compositionality refers to the structural nature of speech which allows a native speaker to project a finite set of *actual* structures (e.g. words) onto a possibly infinite set of *potential* (i.e. new) structures. Currently, the knowledge base parameters are set to a lexicon containing all (actual) syllables of the corpus and to a complete set of phonotactic constraints defining all well–formed (potential) syllables of German (cf. Carson-Berndsen, 1992). However, both these components may be substituted by the relevant syllable lexicon for an English corpus and

[2] Everywhere–predict represents a modification of the PREDICT mechanism of an active chart parser whereby for each standard–predict, in which all the next possible structures are predicted, an initial–predict is carried out in which all possible initial symbols are predicted.

a system of phonotactic constraints defining all potential syllables of English, for example. The knowledge base parameters are thus also responsible for the multilingual aspect of the model. For a particular scenario, this notion can be taken further. Rather than substituting the phonotactic constraints of one language for phonotactic constraints of another, it is also possible to substitute a system of phonotactic constraints which only covers the scenario corpus. This reduces the functionality of the parser as a computational linguistic model but allows for more efficient processing in a product–oriented system by restricting the knowledge base to more specialised data coverage. The approach presented here allows a ranking of various knowledge bases ranging from the most specific (or purely corpus based) to the most general (or complete language) description. The *performance parameters* define the application of constraint relaxation and constraint enhancement, modelling the linguistic performance of the native speaker of the language. The performance parameters allow the linguistic competence constraints to be partitioned into the following performance–oriented categories:

1. Obligatory Performance Constraints (OPCs): feature tiers which must be obligatory in the input, i.e. constraints which may never be relaxed; these define the basis for constraint relaxation.
2. No Contrary Constraints (NCCs): feature tiers whose features must not contradict a particular specification, i.e. constraints which may be relaxed when no contrary information exists; these define the basis for constraint enhancement.
3. Unreliability Performance Constraints (UPCs): feature tiers which can be turned off, i.e. constraints which should be relaxed to the extent that they may be ignored; these define the basis for reducing the importance of unreliable information.

The complete parameter space for these constraints can be empirically verified by iterative tests and evaluation. The optimal parameter settings are defined as those which provide the best recognition results. Constraint enhancement can be applied using the NCCs, which is particularly useful when the input representation is underspecified (noisy). Phonotactic constraints generate a more complete output representation by interpolation as shown in Figure 1. In the case where more information is available in the input than is required by the phonotactic OPCs, this information may be used to specify the output further.

Performance parameters define the possibility of specifying various levels of robustness and allow for a distinction in functionality between optimal native speaker performance, where all constraints would be defined as obligatory, and

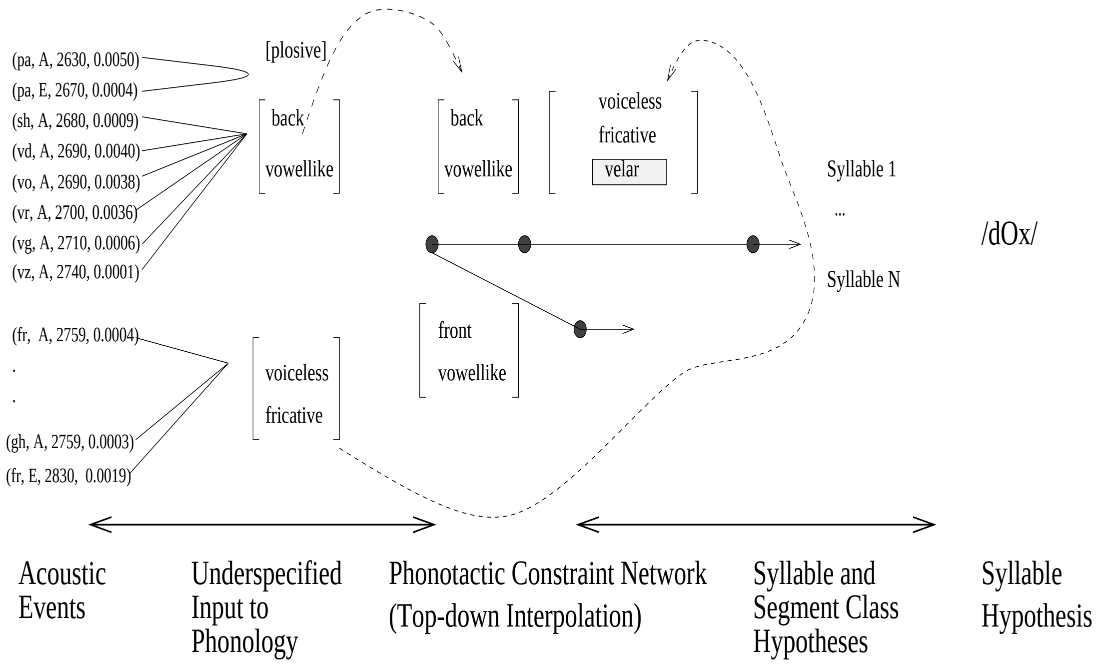

Figure 1: Constraint enhancement.

sub–optimal native speaker performance which is the usual case. The *predict strategy parameters* place constraints on processing within the context of top–down predictions from the word level of processing. A prediction of a word beginning, for example, can be provided by the word parser on the basis of the morphotactic grammar, which is then used during phonological processing to direct and constrain the default *everywhere predict* strategy to perform only an initial–predict or a standard–predict, allowing for variations in predictive power which may correspond to memory limitations and other psycholinguistic factors which play a role in shadowing tasks. The *temporal parameters* are used to define maximum overlap and gap allowances between the events. The upper bound for these parameters corresponds to the maximum length of the phonological word. Those settings for overlap and gap allowances which render the best recognition results provide an indication to the temporal pace and rhythm of the utterance, the temporal aspects of speaker performance at the time of the utterance. However, this issue is part of future research and has not been fully investigated thus far. *Communication parameters* and *system parameters* can be termed non–linguistic, that is to say, they do not influence the specific linguistic aspects addressed in this paper. The communication parameters define the method of communication within the complete spoken language recognition system. The system parameters define the format for the output to the user in

Table 1: Parameter set specifications for the phonological event parser.

Parameter Set	OPCs	NCCs	UPCs	Overlap	Gap
P1	manner	place	{}	50	50
P2	{}	{}	{}	50	50

Table 2: Evaluation results for the phonological event parser.

Data Set	Parametrisation	Phoneme Recognition	Syllable Recognition
A	P1	49.6%	14.5%
A	P2	72.5%	37%
B	P1	54%	21%
B	P2	66.97%	35.19%

terms of data files or screen messages, the extent of verbosity of the parser and a confidence value threshold for individual input hypotheses.

4 Conclusion

In this paper, a parametrisable development environment for phonological event parsing was presented. In particular knowledge base parameters were seen to define the native speaker competence constraints, and multilingual extension and performance parameters were seen to define conditions for the application of constraint relaxation and enhancement. In addition to the role attributed to these mechanisms above, constraint relaxation and constraint enhancement clearly provide the basis for the processing of speech variants since these often do not correspond to the well–formedness constraints of the language. This offers an alternative to the segment–based classification of speech variants described in Carson-Berndsen (1990) whereby the speech variant phenomena are described declaratively in terms of constraints on overlap and precedence which can be relaxed or enhanced during processing. This paper does not claim that spoken language recognisers should be purely knowledge–based. The next step in the development of the phonological event parser is therefore to incorporate stochastic information which will allow for corpus–based fine tuning. Some evaluation results of isolated parameters of the phonological event parser using the German phonotactic knowledge base are presented below based on the diagnostic evaluation procedure described in Carson-Berndsen and Pampel (1994). Two data sets were evaluated: set A consisted of 200 single speaker utterances of read railway information and set B consisted of 82 many speaker utterances of spontaneous scheduling task dialogues.

The results are remarkable for a purely knowledge–based system. It is anticipated that enhancement by stochastic information will lead to further improvement of the recognition results.

References

F. Althoff, J. Carson-Berndsen, G. Drexel, D. Gibbon, K. Hübener, U. Jost, M. Pampel, A. Petzold and V. Strom (1995). Bellex3+1: Linguistische Worterkennung unter Berücksichtigung der Prosodie. Verbmobil Technisches Dokument Nr. 22. University of Bielefeld, University of Bonn, University of Hamburg.

F. Althoff, J. Carson-Berndsen, G. Drexel, D. Gibbon, K. Hübener and M. Pampel (1994). Linguistic word recognition: Bellex3 Manual. Verbmobil Technisches Dokument Nr. 12. University of Bielefeld, University of Hamburg.

S. Bird and E. Klein (1990). Phonological events. *Journal of Linguistics 26* pp. 33–56.

J. Carson-Berndsen (1990). Phonological processing of speech variants. In: *Proceedings of the 13th International Conference on Computational Linguistics (COLING 90)*, volume 3, pp. 21–24, Helsinki.

J. Carson-Berndsen (1992). An event–based phonotactics for German. ASL–TR–29–92/UBI. University of Bielefeld.

J. Carson-Berndsen (1993). *Time Map Phonology and the Projection Problem in Spoken Language Recognition.* Ph.D. thesis, University of Bielefeld.

J. Carson-Berndsen and D. Gibbon (1992). Event relations at the phonetics/phonology interface. In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING 92)*, pp. 1269–1273, Nantes.

J. Carson-Berndsen and M. Pampel (1994). Diagnostic evaluation in linguistic word recognition. Verbmobil Technical Report No. 38. University of Bielefeld.

J. Goldsmith (1976). *Autosegmental Phonology.* Indiana University Linguistics Club, Bloomington Indiana.

K. Hübener (1993). Detektion akustisch-phonetischer Ereignisse. Verbmobil Memo 5. University of Hamburg.

K. Hübener and J. Carson-Berndsen (1994). Phoneme recognition using acoustic events. In: *Proceedings of the 3rd International Conference on Spoken Language Processing*, volume 4, pp. 1919–1922, Yokohama, Japan.

A. Jusek, H. Rautenstrauch, G. Fink, F. Kummert, G. Sagerer, J. Carson-Berndsen and D. Gibbon (1994). Detektion unbekannter Wörter mit Hilfe phonotaktischer Modelle. In: *Mustererkennung 94, 16. DAGM-Symposium Wien*, Berlin. Springer–Verlag.

R. M. Kaplan and M. Kay (1994). Regular models of phonological rule systems. *Computational Linguistics 20* pp. 331–378.

K. Kirchhoff (1995). Two-level modelling of speech variant rules. Verbmobil Technical Report No. 82. University of Bielefeld.

H. Langer (1990). Syntactic normalisation of spontaneous speech. In: *Proceedings of the 13th International Conference on Computational Linguistics (COLING 90)*, volume 3, pp. 180–183, Helsinki.