PROTOTYPICAL SPEECH EVENTS AND SPEECH PERCEPTION

Grzegorz Dogil

Fakultät für Linguistik und Literaturwissenschaft,
Universität Bielefeld, D-4800 Bielefeld 1, BRD.

ABSTRACT

Natural Phonology makes a basic distinction between two process types: the processes which foster the production side of speech (lenitions) and the processes which foster speech perception (fortitions). It has been argued that these two process types are not only functionally distinct but that they also apply to distinct sets of structural positions. Thus the fortitions do not generally apply to the syllable final consonants (more generally speaking - VC - syllabic speech events), whereas lenitions usually spare the syllable initial position (the CV syllabic speech event). From this most basic assumption of Natural Phonology it follows that the CV parts of the syllables should form a group of perceptually salient speech events. In a series of experiments on speech parsing by humans and its simulation by machines we will show that this general prediction of Natural Phonology is strongly supported by phonetic facts.

INTRODUCTION

There are two basic mysteries about natural language: the speed and ease with which it is acquired by a child /*the acquisition mystery*/ and the speed and ease with which it is processed /*the processing mystery*/. The speed and ease of language acquisition is so mysterious because it takes place in the environment of highly deficient input data. There must be then some underlying principles which help children override this deficiency of input data. An attempt to find the underlying principles of the acquisition mystery produced the most important innovations in modern linguistics. The solution of the Natural Phonology - acquisition of phonology as the 'unlearning' (increasing inhibition) of process types - has been not only one of the most original, but also the one with the strongest impact in child phonology.[1] The processing mystery, on the other hand, has been much less popular with the general linguistic community,the natural phonological paradigm included. This paper is an attempt to break with this tradition. Following the basic idea that fortitions foster perception we will illustrate an idea of a parses (a general perceptual mechanism) which considers only these parts of the string where fortitions are allowed.

Similarly to language acquisition, language processing faces a strong input-data-deficiency problem. When we speak we alter a great deal in the idealized phonetic and phonological representations. We delete whole phonems, we radically change allophones, we shift stresses, we break up intonational patters, we insert pauses in the most unexpected places, etc. If to such crippled phonological strings we add all background noise which does not help comprehension either, it is difficult to imagine how the parser is supposed to recognize anything at all. However, even in the most difficult circumstances (foreign accent, loud environment, drunkenness, etc.) we do comprehend speech quickly and efficiently. There must be then some signals in the phonetic string which are particularly easy to grasp and to process. We call these signals *PIVOTS* and parsers working with these signals we call *PIVOT PARSERS*. Until now we have thought only of the phonetic and phonological (or to be more exact - prosodic) pivot parsers, but we believe the idea may be transformed to other types of parsing as well.

THE PHONOLOGICAL *PIVOT*

What are then the pivots in the phonetic string? Dogil [2] argues that at each level of prosodic organization there exist prototypical, unmarked structures which not only manifest themselves in patterns of all natural languages but are also clearly visible in the areas of external evidence such as language acquisition, language loss, and language change. Here we will argue that these ideal prosodic types play an important role in language processing.

At the lowest prosodic level - the level of the syllable - such an ideal type is constituted by a *CV syllable*. That is, the prototypical, unmarked syllable consists of a single consonant followed by a vowel. There is plenty of evidence for this prototype.[2] For example:

- there is no language which does not have CV syllables, but there are many languages which have only CV syllables.
- phonological rules which obliterate syllabic structure usually spare CV syllables.
- CV syllables are acquired first in the process of language acquisition.
- CV syllables are preserved even in the most severe forms of motor aphasia.
- historical syllabic restructuring rules tend towards the creation of CV syllables.
- when subjects are asked to synchronize clicks with syllables it turns out that the clicks are aligned at a point, called the *P-CENTRE* (or 'perceptual centre'), which is close to the CV transitions of the syllable.
- listeners can classify stops by place better than chance when they are given only the first 10-46 msec. of CV syllables.
- the parameters of initial and final transition segments of vowels are not symmetrical in symmetrical syllables (pap, bab, etc.). The parameters of initial transitions may be successfully used as features of the adjacent consonant place of articulation, but the parameters of final transitions are useful as features only in few particular cases.
- when place of articulation cues are different at VC and CV transitions, listeners tend to follow the CV cues.
- speakers try to create temporally more well defined, more precise, articulations near the CV as opposed to VC interface.

All this evidence clearly illustrates the prototypical character of this unit. We claim that this unit is also essential for pre-lexical parsing. *What the parser essentially does is recognize CV syllables in the string.* We propose it does it in the following way:[3]

-- The parser searches for the first CV transition (the 'acoustic boundary' between the consonant and the vowel) and once it has found it, it stops. The parser makes a series of overlapping spectra which spread outwards from the transition point. This gives a *diphonic* representation of the CV syllables.

-- The parser recognizes the syllable. Strictly speaking it recognizes only the unmarked, prototypical CV part of the syllable. These prototypical CV's are stored as diphones in the diphone dictionary. If the syllable contains other units, for example if it is CCVCC syllable (like in the name *Planck*) these other units will be disregarded, and only the CV ([la] of [plaŋk]) will be available after the initial parse.

-- Having identified the syllable the parser makes its first hypothesis about the word of which this syllable is a part.

-- The parsing strategy is continued by jumping to the next transition, i.e. the next CV syllable.

Given all the grammatical, contextual and background knowledge that we possess when parsing strings, the syllabic pivot parser might be actually sufficient for comprehension. Even if it is insufficient in the form that we have presented it so far, it is fast enough to incorporate a number of modification strategies that can make it sufficient for comprehension.

EXPERIMENTAL VALIDATION FOR THE PIVOT: Experiment 1

As a first step we simulated the strategy of the pivot parser by doing some simple speech editing. For the first experiment we chose 10 sentences which show the broad range of syntactic and phonological (segmental and syllabic) complexity to be found in German. The comprehensibility of these sentences had been thoroughly checked on large groups of speakers before the test was devised. These sentences also include the whole range of possible German syllables and a substantial number of consonantal clusters. Speech editing gave us two sets of sentences to be played to the informants. The first set contained the sentences with CV pivots only, and the second set contained the same sentences in which the parts of the pivot were removed. As an example consider the two versions of a sentence *Sie ist nicht leicht zufrieden zu stellen* 'She is not easy to please':

[żi ɩst nɩçt laeçt tsufridən tsu ʃtɛlən] --
unedited version
[zi ɩ nɩ lae tsu ridə tsu tɛlə] --
CV version
[i ɩst içt aeçt uf i ən u ʃ ɛ ən] --
VC version

We randomised the order of these sentences and played each sentence from one set (CV and VC sets, respectively) to 20 native speakers of German and asked them to report on the comprehensibility of the sentences.

The results of this experiment clearly show that the comprehension of sentences of varying syntactic and phonological complexity is significantly better in the case where these sentences are presented in their pivotal - CV form (87.3% correct scores) - than when these pivots are missing from the string (39.8% correct scores).

Furthermore, the syntactic complexity of a sentence does not influence comprehension. Our results show once again that the derivational theory of complexity may not be maintained.

Experiment 2

The hypothesis underlying the Pivot Parser strongly predicts that the CV transitions, which are the only parts of the phonological pivot, are comprehended more quickly and more precisely than these parts of the string which are outside of the pivot, for example the syllable final consonants. This should obviously be the case even if these syllable final consonants *precede* the CV pivot.

In order to check this prediction we asked the informants (the same group as in experiment 1) to tell us the word which they thought was phonetically most similar to the set of non-words. The idea behind this experimental design was that each of the non-words in the stimuli set corresponded phonetically to a minimal pair of existing words, and where one of the members of the pair was similar within the CV pivot and the other pair member was similar outside of the pivot, but the similarity point al-

[1] Cf. Edwards [6], Dressler [4].

[2] Cf. Clements & Keyser ([1], 19-23, 28ff.);Ohala & Kawasaki ([10], 115-119); Kelso, Saltzman & Tuller ([9], 50ff.); Dogil & Braun [3] for the most recent treatments of CV's status in phonetics and phonology.

[3] The presentation here is a rough outline of the Pivot parsers recognition strategy. More detailed examples will be provided in the section 'The PIVOT as word recognizer'.

ways preceded the CV pivot. Table I gives you some of the examples from this non-word set together with the minimal pair set, which we tried to elicit, and the most frequent responses of the informants.

TABLE I

Non-word stimuli	Words (expected results)	Most frequent results
ELDE	ENDE - ELFE - ENTE	ENDE
WEPTE	WESPE - WESTE	WESTE
ALVEN	ALGEN - ALBEN	MALVEN
MAFTEL	MANTEL - MANDEL	MANTEL
WEKPE	WESPE - WESTE	WESPE
ELPE	ELFE - ELCHE	ELBE
RAUSCHPE	RAUSCHTE - RAUCHTE	RAUPE
RESKE	RESTE - RECHTE	FRESKE
WÄRLER	WÄRTER - WÄCHTER	WÄHLER
GÄNCHE	GÄNSE - GEMSE	KÄNNCHEN
RAULTE	RAUCHTE - RAUSCHTE	RAUTE
MANSEL	MANTEL - MANDEL	MAMSEL

If the informants chose the similarity within the CV pivot - for example, if they reacted with ENDE (end) to the stimulus ELDE - this would mean that they were using the pivotal recognition strategy rather than the left-to-right strategy. If latter were the case they should react with ELFE (elves) to the stimulus ELDE.

A tendency which was noticed was that the informants were not really trying to match the non-word with the words in which one CV pivot was identical, but showed strong preference for choices where both CV's matched. For example, the stimulus RESKE, which we expected to activate such words as RESTE or RECHTE, actually activated a word FRESKE in which both CV pivots match the stimulus. The same was the case with the non-word stimulus WÄRLER, which did not activate either the word WÄRTER (predicted by the left-to-right matching strategy) nor the distant word WÄCHTER, but rather the word WÄHLER which almost matches the non-word at both pivots. The only difference between the non-word WÄRLER as it was pronounced by the instructor and the response WÄHLER was in the length of the vowel, property which is not distinctive in german phonology. A similar case holds for numerous other pairs like: GÄNCHE - KÄNNCHEN, RAULTE - RAUTE, RAUSCHPE - RAUPE, MANSEL - MAMSEL.

THE PIVOT AS WORD RECOGNIZER

The PIVOT PARSER predicts that some parts of the string - the CV-pivots - are perceived more precisely and more exactly than other, non-pivotal, parts of the string. How is this *CV parser* supposed to work? We mentioned some general principles in the third section of this paper (The phonological PIVOT). The details will be illustrated immediately below.

Let us suppose the CV parser is confronted with the word *donkey* [doŋkι]. As a first step the speech envelope of this word will be stored in the form given in the figure below.

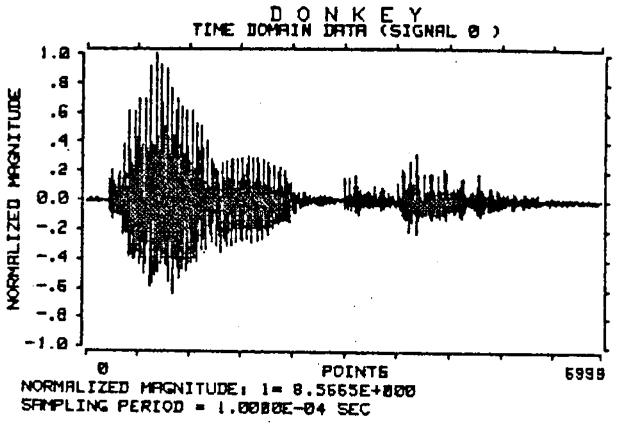


FIGURE 1: Speech envelope of the word *donkey*.

The signal will be transformed into its intensity tracing. This intensity tracing is the input to the segmentation algorithm. As we already mentioned in the section 'The phonological PIVOT' our aim is to sample acoustic information at a point of a transition between a consonant and a vowel within a CV. Stevens, in a series of experiments on acoustic cue recognition (Stevens, [11]), has provided convincing evidence for the perceptual importance of acoustic events in the vicinity of the 'consonant-vowel boundaries'.

> ...these brief time intervals when there is a rapid change in spectrum or amplitude create regions that are rich in information concerning the phonetic features in an utterance. (...) it would appear that a great deal of information is carried by these one-eighth-inch time slots in the spectrogram - much more than one would expect on the basis of the space they occupy in linear time.
>
> (Stevens, [11]: 253)

Moreover, these 'consonant-vowel boundaries' are relatively well marked by the speech producing system. At places where they occur there is usually an abrupt change in the amplitude. This change has been often considered (and used) as a cue to a boundary between individual speech sounds within an utterance. Our approach to these regions of abrupt amplitude change is quite different. We consider them as landmarks of a segmenting algorithm which considers them as *centers* (pivots) of units to be used in speech recognition.

Such an algorithm is being developed at the University of Bielefeld by Dafydd Gibbon (Gibbon, [8]). Given the intensity tracing as in the figure above, it automatically fixes these points where the most abrupt changes occur. Obviously, we are interested only in these changes where the amplitude jumps (characteristic of CV) not where it makes a dip (this is the characterization of the VC transition). Having fixed the first CV transition region we start sampling spectral information in its vicinity. We fix the Hamming window of the length of about 20 msec. and center it around the transition area. We suggest that the mechanism which samples spectral information should never leave the transition region. The only method which guarantees this is to make stepwise growing spectra with the transition

point (the bar of the segmenting algorithm) as the center. The figure below illustrates this method.

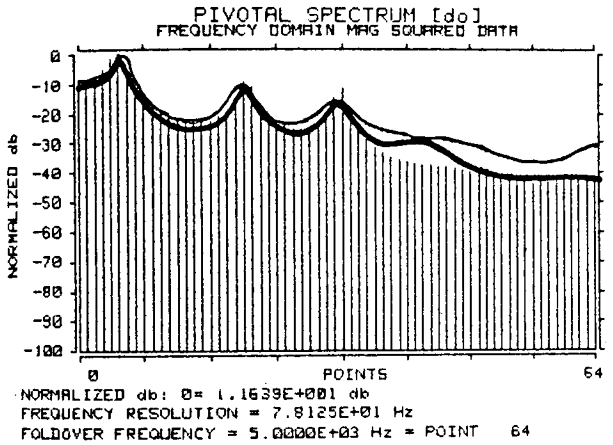
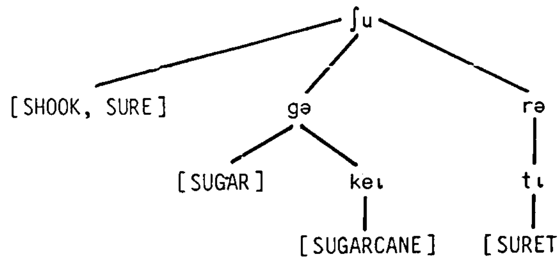


FIGURE 2: Pivotal spectra selected around the transition from [d] to [o] in *donkey*. The bars indicate the central spectrum, the thick line the next largest one, and the thin line the largest one.

These 'pivotal spectra' will be the input to the speech decoder. The difference to the usual procedure is that the acoustic information from the transition area is always present within the spectrum. What changes is only the amount of information which is sampled in the immediate (left and right) neighbourhood of the transition. We believe this change of perspective to be important, particularly from the point of view of 'time normalization' in speech recognition systems. It might turn out that spectra corresponding to different time-size windows around the transition correlate with various speech rates. If this were the case, we would have had a mechanism of encoding various speech rates in the spectral matrix itself.

The [do] of *donkey* has been, thus, decoded, and it is forwarded to the lexical recognition procedure. The lexicon on which this procedure will be stimulated is the 20.000 word phonetic lexicon of American English (Pocket Merriam-Webster, cf. Zue [12]). We have coded it in such a way that each lexeme is stored as the series of CV syllables that it contains. The lexical search itself is taken care of by a programme in PROLOG. This programme searches up lexical trees, which are organized in such a way, that each non-terminal node contains a CV syllable, and each terminal node contains word (or words) which can be made up of this CV syllable, and other CV syllables combinable with it. Such a tree is much easier to illustrate than to describe (after all illustration is what trees are for). The tree in the figure below illustrates the words starting with the CV syllable [ʃu].

FIGURE 3: Lexical subtree for words containing [ʃu] as their first CV syllable.



To return to the recognition of the word *donkey* - we have decoded the first CV [do]. The Pivot parser in PROLOG will give us a tree with this CV at the top, and all possible CV's which may be combined with it will be its daughters on the tree. All in all, they form a cohort of 45 words. However, as soon as the speech decoder decodes the second CV syllable of *donkey* - the [kι] - only one word remains: DONKEY.

This recognition procedure is very fast, and the reduction of initially large cohorts is quite optimal. It seems to be the case that the CV syllables do not combine so freely to form words as one would imagine they should. We tried out this recognition procedure on a number of words, and we never got really bad results. Consider the well-ploughed example *trespass*. After decoding the first of its CV syllables [trɛ] we are confronted with 21 word candidates, but having decoded the [pa], we immediately recognize *trespass*. Even in complex cases, where the division of the string into the CV syllables is difficult, and where there are many other consonants between the pivots, recognition is very fast, and, actually, unique. The word *abstract* [æbstrækt] is such an example. Out of its five consonants only one is decoded by the parser - the [tr] of the second CV syllable [træ].[4] Still the blank CV's - [æ] and [træ] - suffice to reduce the cohort into the following words - ABS'TRACK, 'ABSTRACT, ABSTRACTS and ABSTRACTION!

We have shown here that the parsing strategy of the PIVOT, when applied to words as heard in isolation, enables very fast and efficient recognition. This is obviously true mainly of polysyllabic words, the monosyllabics are a problem. Who would, however, want to stop half-way and consider recognition of words spoken in isolation?! The real test for any model of speech parsing is the recognition of connected speech. Let us see what the PIVOT has to offer in this area.

THE PIVOT AS PARSER OF CONNECTED SPEECH

We decided to make use of the apparently limited 'combinability' of CV syllables by giving the parser not just words, but the whole utterances in their PIVOTAL - CV form. Actually, PROLOG's command for this subroutine is - *get_sentence*. Thus, when we fed our parser with the string like the following:

```
get_sentence ([wɔ, də, ðι, pι, və, du], X).
```

[4] Note that [tr] is a monosegmental affricate.

one sentence, with variation at two structural positions was our result. Incidentally (accidentally), the first of the 'possible' variants is the sentence that we were aiming at: *WHAT DOES THIS PIVOT DO.* Note, the combination of six CV syllables was analysed into one, single sentence with only slight variation in two positions. In the input we skipped a number of consonants (codas), we did not mark any boundaries between words, and we did not use any repair strategy - neither syntactic nor intonational nor semantic, nor frequency of cooccurrence. It was just the CV PIVOTs which were matched with the lexicon! If you consider the size of the lexicon (20.000 words), this result clearly speaks for the fact that the PIVOT is not the worst of the connected speech recognizers. It is also not one of the slowest! Although PROLOG is not the fastest of the 'intelligent' languages and although the machine on which it is implemented, does not do any MIPS, the simulations described here are all a matter of milliseconds. We tried a couple of other sentences. We will not give you them all, but the consideration of the modest one - *THIS PIVOT SIMULATES HUMAN PERCEPTION* -. will give you the idea where the problems lie. The PIVOTAL input form for this sentence is : [ðɩ, pɩ, və, sɩ, mju, leɩ, hju, mə, pə:, sɛ, ʃə]. As an output we got 16 sentence analyses and the last one[5] was as the following:

[[THIS] [PIVOT] [SIMULATE] [HUMAN, humid, humus] [PERCEPTION]]

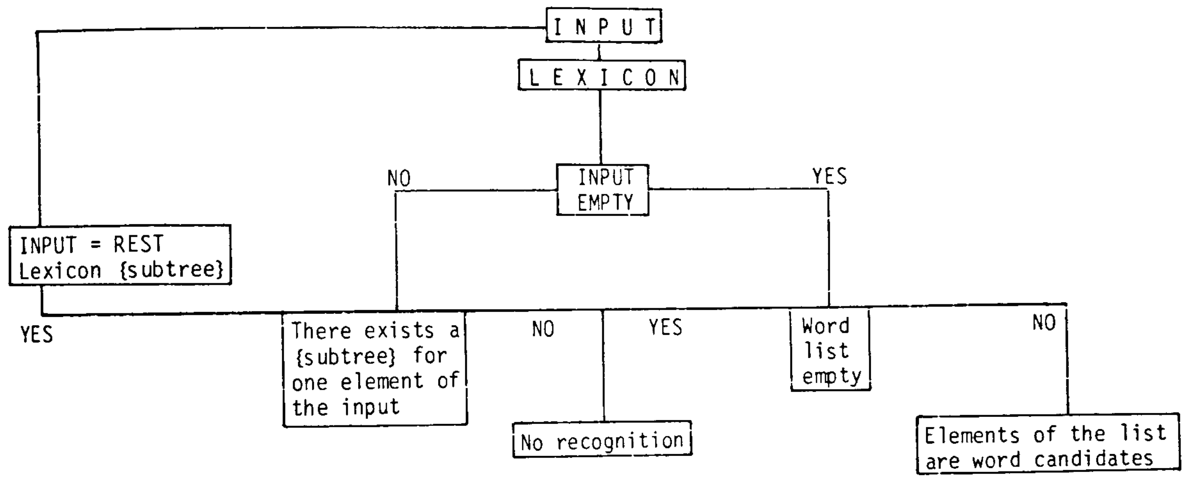
As you see, we do not need some sort of top-end parsing to get the optimal reading out of this output (i.e. to eliminate HUMID and HUMUS as possible word candidates). A quick look at other 15 alternatives makes it also apparent that some syntactic parser would be of great help in eliminating most of these analyses. However, the work on parsers for syntax, semantics and other 'higher' knowledge systems has advanced so much, that we do not doubt that they can help us. What is much more important is, that in this, and in every other case of simulated speech recognition which we have carried out, it was always the case that all the words in an utterance were recognized in at least one of the output analyses.

This gives us a guarantee that the bottom-up PIVOT speech decoder and recognizer may be considered to constitute a fast, efficient and *sufficient* input to the top-end parsing strategies. As far as we can see, it is the optimal 'ear' for speech recognition.

In summary, and for those of our readers who appreciate pictures more than text, we give a graphic sketch of the lexical recognition procedure which we tried to describe in words in this paper.

[5] The PIVOT, as it is implemented now, tries to build *shortest* possible words as soon as the allowable CV combination has been found. The rest of the possibilities are found by backtracking. This need not actually be the optimal solution. In our example under discussion the *longest* allowable CV combination leads to the best results. The problem as to which of the combinations should be analysed first, is an empirical problem, which can be satisfactorily answered only after numerous simultations and psycholinguistic experiments have been carried out and analysed. At any rate, the PIVOT belongs to the 'no alignment' class of recognition theories. These theories do not pretend to 'know' where the boundaries are in the signal, but they single out some speech events (e.g. CV or a distinctive feature) and allow it to combine with any other speech event of the same time. The boundaries arise through constraints on the combinability of these units (CV's in our case).

TABLE II



Input :	List of CV's
Lexicon :	{subtrees}
Subtrees :	[CV, WORD LIST, SUBTREES*] OR [ ]
Word list :	[(WORD*)]

CONCLUSION

The most general conclusion of the PIVOT PARSER, and the one which makes our research programme distinct from all other approaches to language processing, is, that our parser does not require the exhaustive processing of strings, and that it explicitly claims that all language processing is based firstly and foremostly on the prototypical, unmarked units, which we called PIVOTS. Wheather our choice of the CV as the phonological PIVOT (the prototypical phonetic gesture) is correct or not, is, given the plausibility of this most general conclusion, only of secondary importance. However, the strong support that CV gets from the work done within the theory of Natural Phonology is an additional argument to consider it a prototypical speech event.

Although it is true that the event theory in phonetics is just at its beginnings (cf. Fowler [7]), and the event theory of phonology is only emerging (out of some ideas in the Natural Process Phonology, cf. Dressler [5]), the enterprise of replacing the segment oriented approach with an event oriented approach may prove highly rewarding in the studies of speech.

REFERENCES

[1] Clements, G.N. & S.J. Keyser (1983). *CV phonology*, Cambridge, MA., The MIT Press.

[2] Dogil, G. (1985). *Theory of Markedness in nonlinear phonology*, Habilitationsschrift, University of Bielefeld. (Available from the author)

[3] Dogil, G. & G. Braun (1986). *The Pivot Model of Speech Parsing*, distributed by LAUD (Linguistic Association University of Duisburg)

[4] Dressler, W.U. (1984). Explaining Natural Phonoloy, *Phonology Yearbook* 1, 29-53.

[5] Dressler, W.U. (1985). *Morphonology: the dynamics of derivation*, Ann Arbor, Karoma Press.

[6] Edwards, M.L. & L.D. Shriberg (1983). *Phonology*, San Diego, College Hill Press.

[7] Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective, *Journal of Phonetics* 14, 2-28.

[8] Gibbon, D. (1986). Prosodic parsing with parallel sequence and hierarchy incrementation (PSI/PHI), ZiF Workshop on Speech Parsing, October 15-17, Center for Interdisciplinary Research (ZiF), Bielefeld.

[9] Kelso, J.A.S., Saltzman, E.L. & B. Tuller (1986). The dynamical perspective on speech production: data and theory, *Journal of Phonetics* 14, 29-59.

[10] Ohala, J. & H. Kawasaki (1984). Prosodic phonology and phonetics, *Phonology Yearbook* 1, 113-129.

[11] Stevens, K.N. (1985). Evidence for the role of acoustic boundaries in the perception of speech sounds, in V. Fromkin (ed.), *Phonetic linguistics: Essays in honour of Peter Ladefoged*, New York, Academic Press, 243-257.

[12] Zue, V. (1983). Proposal for an isolated-word recognition system basea on phonetic knowledge and structural constraints, in Cohen & Broecke (eds.), *Abstracts of the Tenth International Congress of Phonetic Sciences*, Dordrecht, Foris, 299-307.