THE PIVOT PARSER

Grzegorz Dogil

ABSTRACT

Similarly to language acquisition, language processing faces a strong input-data-deficiency problem. When we speak we alter a great deal in the idealised phonetic and phonological representations. We delete whole phonemes, we radically change allophones, we shift stresses, we break up intonational patterns, 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 recognise 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.

THE PHONOLOGICAL PIVOT

What are then the pivots in the phonetic string? Dogil (1987) 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 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. 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 synchronise 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.

* Universitä t Bielefeld, Fakultät für Linguistik und Literaturwissenschaft.
- 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 recognise CV syllables in the string. Let us suppose the CV parser is confronted with the word donkey [daNki]. As a first step the speech envelope of this word will be digitalised. The signal will be transformed into its intensity tracing. This intensity tracing is the input to the segmentation algorithm. Stevens, in a series of experiments on acoustic cue recognition (Stevens, 1985), 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, 1985: 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.

Having fixed the first CV transition region we start sampling spectral information in its vicinity. We fix the window of the length of 20 msec. and center it around the transition area. That is, the resulting spectrum will include acoustic information from the consonant release (about 10 msec. of it) and from the vowel. This will give us a spectrum of the CV transition in the syllable [da] of [daNkI]. 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.

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.

This [da] of donkey has been, thus, decoded, and it is forwarded to the lexical recognition procedure. The lexicon on which this procedure will be simulated is the 20,000 word phonetic lexicon of American English (Pocket Merriam-Webster). 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 short programme in PROLOG. To return to the recognition of the word donkey - we have decoded the first CV [da]. The Pivot parser in Prolog will give us a tree with this CV at the top, and all possible CVs 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 the speech decoder decodes the second CV syllable of donkey - the [ki] - 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 our well-ploughed example trespass. After decoding the first of its CV syllables [trE] we are confronted with 21 word candidates, but having decoded the [p@], 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 [@br@kt] is such an example. Out of its five consonants only one is decoded by the parser - the [tr] of the second CV syllable [tr@]. Still the blank CV's - [@] and [tr@] - suffice to reduce the cohort into the following words - 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 sting like the following:

get_sentence([Ha, d@, DI, pI, v@, du],X).

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 not semantic, nor frequency
of cooccurrence. It was just the CV PIVOTs which were matched with the lexicon! In order not to become over enthusiastic, 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: [DI, pI, v@, sI, mju, le/, hju, m@, p@:, sE, S@]. As an output we got 16 sentence analyses, and the last one was as the following:

[[THIS] [PIVOT] [SIMULATE] [HUMAN,HUMID,HUMUS] [PERCEPTION]]

As you see, we do 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). 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 recognised in at least one of the output analyses. This gives us a guarantee that the bottom-up PIVOT speech decoder and recogniser 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.

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.

REFERENCES