Pseudo-Articulatory Representations and the Recognition of Syllable Patterns in Speech.

William H Edmondson & Li Zhang

School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK.
W.H.Edmondson@cs.bham.ac.uk,L.Zhang@cs.bham.ac.uk

Abstract

This paper presents an account of syllable structure as the basis for organizing articulatory activity. This contrasts with the serial organization of more conventional phonetic segments. It is demonstrated that working with syllables in this way can provide the basis for linguistically motivated speech recognition using the previously reported notion of the Pseudo-Articulatory Representation (PAR).

1. Introduction

Previous work [1, 2] has established the notion of Pseudo-Articulatory Representation (PAR). Briefly, the PAR is based on the linguist's conception of binary distinctive features – a set of parameters which are both language and speaker independent, and which categorize speech sounds in terms of a very abstract and a temporal model of the vocal tract. The PAR takes this abstract model and injects realism in two ways – by making the feature values continuous where appropriate, instead of binary, and by providing values as a continuous function of time, instead of segment by segment.

Iles [3] established that speech could be recognized on the basis of recovering PAR values from the speech waveform. He did this by developing complex formulae which provided equivalencies between formant values recovered from speech signals, and PAR values. The demonstration did not yield textual representations, but showed that the recovered PAR values were valid by using them to regenerate speech using a PAR controlled synthesizer. Iskra [4] looked specifically at the use of PARs to tackle the speech recognition problem, following Iles, but was unable to develop his approach much further because of the difficulty in recovering articulatory information irrespective of any supposed segment labels.

In this paper we show how a model of syllable articulation can be used with PARs to provide a general articulatory transcription of speech without phonetic labelling. This will form the basis of a speech recognition system.

2. The syllable

There is a long established debate on the relative merits of the syllable and the segment as the basic unit of articulation. Bell and Hooper [5] note that discussion of sonority as an organizing principle for syllable structure goes back to the late 19th century. More recently Kaye [6] has argued that incorporating syllable structure into phonological representations brings benefits, and rather dramatically he has also argued that "the phoneme is dead" as a concept of phonological interest. In this paper we assume that the syllable can be accepted as a unit or domain for organizing articulatory activity, and we explore the idea that it is the right unit when considering speech recognition processing.

2.1. Structure of the syllable

There are several different ways of analyzing the syllable, and our first question is which is most useful as the basis for work on automated speech recognition? Conventionally, speech segments are considered to be articulatory units, and these are organized in sequences which are patterned as syllables. In this way syllables are analyzed in terms of sequences of consonants and vowels: V, CV, CVC, CCVC, and so forth. If this model is to be useful in speech recognition, the consonants and vowels must be recognized first, and then their patterning as syllables analyzed to provide structural constraints. Whilst this can assist the recognition process, that process begins with identification of candidate consonants and vowels, a step we seek to avoid (on the grounds that it assumes too much about the articulatory organization of speech; in any case the poor accuracy of such recognition is part of the problem we are trying to solve).

Syllables can be analyzed as larger units with structure, and there are two candidates for this. The most widely accepted model is of the syllable as Onset+Rhyme (sequentially) with Rhyme being Nucleus+Coda. This is shown below in Figure 1.

The onset and the coda are not always present in every syllable. The three elements are not segments in
the conventional sense – for example the onset can be a cluster of consonants. The nucleus is not always a vowel – as in the second syllable of the word ‘button’.

This analysis is thus more general or abstract than the conventional CVC type of sequence, but it offers little more than that when it comes to recognition systems – one can use the abstract structure as an organizing constraint (‘maximize the onset’ and so forth) but recognition must be still attempted independently of the syllable.

A different way of working with the syllable as a unit is to use sonority as the organizing principle. The scheme here is to note, as others have done before [5], that syllables are ‘sonority waves’. The sonority of the speech sound builds up during the onset, to the peak value at the nucleus, and drops away again in the coda, the whole cycle repeating as syllables are produced in sequence. In this model it is envisaged that individual speech sounds/segments have sonority values (on a scale of perhaps 1-10), and thus the constraints on sequential arrangements of consonants in the onset and the coda are explained in terms of sonority contours. This provides additional constraints when considered in comparison with CVC type models of syllable structure, and this can assist recognition.

### 2.2. Articulatory pattern in the syllable

The approach we have taken focuses instead on the notion that a syllable is basically an articulatory unit. We have chosen to describe this, rather abstractly, as follows:

```
transition    syllabic target    transition
```

This expands to a more layered structure, shown in Figure 2, giving three layers altogether, where ‘s-tar’ means syllable target, ‘d-tar’ means dynamic target, ‘tr-target’ means transition target, ‘tr’ means transition. The use of bold font in Figure 3 means that the identified component is marked for a specific ‘phonetic’ value, normal font means that the component is not identified as marked (it may have a complex specification, or no specification), italic means the component cannot be marked. Clearly, s-tar is always marked in reality (else there would be no syllable).

Figure 1

![Figure 1](image)

In this scheme articulatory activity must consist of tr, x-tar, tr, x-tar, tr, x-tar etc. where syllable nuclei are marked by x = s, and where phonetically irrelevant tr are tr. Typically, then, a CCCVCCC syllable might look like:

```
tr, tr-tar, tr, d-tar, tr, tr-tar, tr, s-tar, tr, tr-tar, tr, d-tar, tr, tr-tar, tr
```

An example of how this might be used for the English word ‘apt’, is shown in Figure 3.

```
tr, s-tar, tr, tr-target, tr, d-tar, tr, tr-target, tr
[æ]            [>]p]            [pt]             [t<]
```

Figure 3

This shows that the articulatory detail can be labelled ‘phonetically’ but this does not equate to phones. The [p] is shown not as a phone, but rather just as the closure phase; likewise the [t] is shown as release phase. Additionally, complex articulatory activity, without phonetic significance but required for the phonetic string in which it is embedded, can be recorded, as in the case of the change in point of obstruction in the phase, or component, labelled ‘d-tar’ above.

### 3. Use of PARs

We now show how the details of syllable articulation can be recovered from Pseudo-Articulatory Representations.

We chose to work with idealized PARs because we want to determine the feasibility of relating PARs to syllable structure without any additional problems which might arise from the use of PARs computationally derived from speech signals. If we can
The top section shows the spectrogram of the utterance “There is usually a value.”. 
The middle 4 traces show the idealized feature trajectories of high, back, round, tense. 
The bottom section shows in schematic form the recovered syllable positions.
demonstrate the feasibility of the relationship, we will go on to consider the problems of computationally derived PARs.

The idealized PARs were produced by ascribing four feature values to every segment in the transcription files. The values for vowels were taken from the vowel model, used by Iskra [4]. The values for consonants were taken from the consonant model [4]. Smoothed transitions between ideal targets are presented, as well as the targets themselves. Between targets there is a significant change in the feature values. For any idealized target, especially vowel targets, the trajectories remain stable, and thus the feature values as well. By using the articulatory pattern in the syllable, which we have discussed in 2.2, as a rule, an algorithm has been created to identify the targets and transitions in the utterance context. For example, at the beginning of the utterance, after the first transition, there will be a target. It has an uncertain specification because in the syllable onset there can be more than one consonant or no consonant at all. The algorithm will read following data points along the sequences of feature values to recover further information. On the basis of evidence from the following data, the unknown articulatory activity can be marked for a specific articulatory value. The subsequent articulatory activities are marked in the same way, using data even further down the sequences as well as information from the already labelled articulatory activities. In this way the syllable structures are recovered in sequence. Meaningful syllable structures for one utterance have been derived in this way, and are shown diagrammatically in figure 4 and in detail in figure 5.

![Figure 5 the analysis results between 0.7s and 0.8s by every 10ms](image)

The algorithm seems promising although currently it is based on idealized PARs.

4. Conclusions

Speech processing for recognition is conventionally concerned to recover a string of phones from the acoustic waveform. We have chosen here to explore the idea that it might be easier to recover strings of phonetically unlabelled syllables, and to use this information to recover phonetic detail without requiring that this detail be expressed in terms of phones.

Our approach has been to consider idealized Pseudo-Articulatory trajectories as the basis for recovery of detail in a simple model of syllabic articulatory patterning. Working with a limited data set, at the moment, we have shown that it is in fact possible to recover the desired details without resorting to models of the phone, or to models of the syllable as a sequence of phones. This suggest that the syllable is the right articulatory unit for speech recognition processing.

The next step in our work will be to attempt syllable recovery using computationally derived PARs, in the manner of Iles [3], and this will be followed by attempts to label the various components of the syllables with enough phonetic detail to permit recovery of the linguistic representation. It remains to be seen whether or not phonotactic constraints, or patterns based on sonority contours, will also be required to assist with the labelling of the syllables. Ultimately, phonemic labelling and morphological recognition must underpin the recognition process, and we consider this will be supported by syllable identification.

5. References


