SPEECH RECOGNITION USING SYLLABLE PATTERNS

Li Zhang & William H Edmondson

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

ABSTRACT
This paper presents an account of the use of syllable structure as the basis for a novel approach to speech recognition. This contrasts with the serial organization of more conventional phonetic segments, and their use in speech recognition systems. It is demonstrated that working with syllables provides the basis for linguistically motivated speech recognition using the previously reported notion of the Pseudo-Articulatory Representation (PAR). The results are very promising taking into account the preliminary nature of the work and the novelty of the approach. A related paper [1] deals with theoretical issues in greater depth.

1. INTRODUCTION
Earlier work [2,3] 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 atemporal 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.

In previous work [4], Iskra looked specifically at the use of PARs to tackle the speech recognition problem, following the work of Iles and Edmondson [5], but was unable to develop the 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
The rôle of the syllable in speech is complex. It is both segmental and supra-segmental, in conventional terms, and in speech processing for recognition syllables can contribute in both domains. Syllables vary in weight or complexity, and pattern in many languages in ways that support or enable prosodic structuring in speech – within words and across groups of words. Clearly, speech recognition systems can exploit such prosodic patterning, both as part of the general enterprise and as a specific facet of word recognition. This aspect of the syllable – its value as some sort of segment in larger structures – does not concern us here.

2.1. Structure of the syllable
The internal structure of the syllable can be described or analyzed in various ways (and we include later a scheme first reported in [6]). Most obviously, syllables can be viewed as groupings of segments – a vowel nucleus with up to three consonants before and after. More precisely, we can write the possibilities as $(C)(C)(C)V(C)(C)(C)$ where the brackets show the consonants to be optional. Phonotactic constraints which are language specific limit the sequences of consonants (e.g. see [7] for details for English).

Another conventional approach to syllable structure is shown in Figure 1, where the tree structure defines a frame for any potential syllable.

```
syllable
   /\   /
  /   \ /
onset rhyme
```

Figure 1

The syllable structure models can be quite complex, with specifications for components of both onset and coda (cf. [7]). Note also that onset and coda may not be evidenced, for example when the syllable is just an isolated vowel, and that the structural approach does not specify that the nucleus must in fact be a vowel.

The analyses of linguistic data can also get complicated, for example when geminate consonants are involved, but the outline given here illustrates the point that syllable structure is more than just the sequential arrangement of vowels and consonants.

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 that syllables are ‘sonority waves’ [7,8]. 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-tar’ means transition target, ‘tr’ means transition.

\[
\begin{array}{c}
\text{tr} & s\text{-tar} & \text{tr} \\
\text{tr} & d\text{-tar} & \text{tr} \\
\text{tr} & \text{tr-tar} & \text{tr} \\
\end{array}
\]

\[
\begin{array}{c}
\text{tr} & \text{d-tar} & \text{tr} \\
\text{tr} & \text{tr-tar} & \text{tr} \\
\text{tr} & \text{tr-tar} & \text{tr} \\
\end{array}
\]

Figure 2

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).

In this scheme articulatory activity must consist of tr, x-tar, tr, x-tar etc. where syllable nuclei are marked by x = s, and where phonetically irrelevant tr are tr. Typically, and assuming that the targets correspond with phonetic segments identifiable as consonants (see the related paper [1] for discussion) a CCCVCCC syllable might look like:

\[
\begin{array}{c}
\text{tr} & \text{tr-tar} & \text{tr} & \text{d-tar} & \text{tr} & \text{tr-tar} & \text{tr} \\
\end{array}
\]

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

\[
\begin{array}{c}
\text{tr}, & s\text{-tar}, & \text{tr}; & \text{tr-tar}, & \text{tr}, & \text{d-tar}, & \text{tr}, & \text{tr-tar}, & \text{tr} \\
\end{array}
\]

[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.

Before moving on to discuss how the syllable structure information might be used in speech recognition it is necessary to remind the reader how Pseudo-Articulatory Representations are derived.

3. Mapping procedure
The first step in the recognition processing is the establishment of a mapping between PARs and acoustic parameters. This is done in order to set the system up for subsequent processing – it is not undertaken for ease of mapping.

Formant frequencies, amplitudes and bandwidths are chosen as acoustic parameters. (The work has also been done using cepstral coefficients, but we do not report this here.) The speech data are obtained from the TIMIT database and for the time being only one speaker is taken into account. For each phone, 18 formant parameters are calculated based on all the available occurrences of this phone.

3.1. Vowel model
The mapping is done for vowels to start with. In the previous work [2,3], the PAR description is obtained by selecting four features: high, back, round, tense and ascribing a value between 0 and 100 to every vowel based on the data provided by Ladefoged [9]. In the current work a new PAR description has also been used. Depending on the position of the tongue, whether it’s in the front or the back of the mouth, up or down, vowels can be represented as located on a quadrilateral; i.e the Cardinal Vowel System. We adapt the vowel quadrilateral by using the lower left corner as the origin of a polar coordinate system. Then we use ‘r’ (called ‘distance’ below) as a measure of the degree of constriction and ‘θ’ (called ‘theta’ below) as a measure of the vocal tract length. The new description is obtained by replacing the features: high and back with the new features: theta and distance. Finally, the new PAR description is obtained by selecting four features: theta, distance, round, tense and ascribing a value between 0 and 120 for theta, a value between 0 and 150 for distance.

3.2. PAR derivation for consonants
In the earlier work the consonants are located on the vowel quadrilateral – or more precisely, just outside the periphery. In order to determine new PAR values for consonants the diagram suggested by Ball [10], which is closer to physical reality, is adopted. As in the previous work this permits the account of vowels to be extended to cover consonants (in terms of target configurations, but not dynamic aspects). Our interest in Ball’s proposal is that it offers the possibility that the acoustic parameters recovered from speech can be more easily linked to the palatal, velar, uvular and pharyngeal places of articulation, and thus to articulatory behaviour and phonetic categories.
4. RECOGNITION

In the recognition process three successive stages can be clearly distinguished. The first stage is responsible for the transition from the acoustic representation of the incoming signal to the pseudo-articulatory representation with feature trajectories available as a function of time. The second stage concerns the movement from the pseudo-articulatory representations to the recovered syllable structures and produces a sequence of the recovered syllables. The third stage focuses on the transition from the syllable patterns to the phonetic level of description and produces a sequence of phone labels. This third stage can be augmented by other phonetic labeling data derived using conventional techniques.

4.1. Transition from the acoustic to the Pseudo-articulatory level

The first stage of the recognition process establishes every 10 msec a set of 18 formant parameters for the incoming speech. A brute search mechanism is used which by gradually reducing the solution space determines four PAR values for each set of 18 formant parameters. As a result of this, an utterance is described with both a set of values for theta, distance, round, tense every 10 msec and a set of values for high, back, round, tense every 10 msec. When plotted, these values present feature trajectories for that utterance.

4.1.1. Evaluation by resynthesis

As a result of the brute search mechanism two series of four pseudo-articulatory values are produced for each 10ms of speech in the test file. This is repeated for a set of 10 sentences. The results are plotted as trajectories for respective features and compared to the idealized ones. The idealized trajectories are produced by ascribing four feature values (obtained from tables in textbooks) to every segment in the transcription files. And as noted above in 3.1 two different sets of four idealized feature trajectories are produced. The values for vowels are taken from the vowel model. The values for consonants are taken from the consonant model.

It is hard to create a general picture of how close the recovered trajectories are to the idealized ones, though the idealized trajectories seem to be a reasonable approximation of the new ones, at least on average, since the recovered trajectories clearly contain considerably more peaks and troughs. But another way of evaluating recognised trajectories is resynthesis. A conventional synthesis procedure is used to do the resynthesis work. The quality of the synthesised speech is evaluated by listening to it. All the sentences are comprehensible and clear, and sound natural. The only differences with respect to the original sentences are a few clicks which might have been caused by some inaccuracies in file handling.

4.1.2. Smoothing the computationally derived PARs

The recognition procedure focuses next on such aspects as smoothing the computationally derived PARs, because the recovered trajectories clearly contain too many peaks and troughs. An averaging algorithm has been used to smooth the trajectories, and the synthesised speech creates again. All the sentences are comprehensible and clear, and sound natural too.

At this point, we consider the PAR data are good quality and suitable for running the syllable recovery algorithm.

4.2. Syllable recovery

Previous work [6] has considered idealized PAR trajectories as the basis for the syllable recovery. Here we are using smoothed PAR trajectories recovered from speech. The next step in our account is to demonstrate how the details of syllable articulation can be recovered.

In the smoothed trajectories, smoothed transitions between smoothed targets are presented, as well as the targets themselves. Between targets there is a significant change in the feature values. For any smoothed 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 known 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.

Finally, a sequence of tr, x-tar, tr, x-tar, etc, has been derived as well as syllable positions. As we have discussed in section 2.1, 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. Since we know the sequence of the articulatory events and the positions of the syllables, an algorithm is used to find the syllable peaks according to the principle mentioned above. Finally, the ‘sonority waves’ are produced in sequence and are shown diagrammatically in figure 4. The sentence is: There is usually a valve.

4.3. Finding a phone sequence

At this point, the recovered syllable patterns are used to label the various components of the syllables in order to find the best matching sequence of candidate phones by calculating the distance between each set of four incoming feature values and the idealized values used in initial construction of PARs. At each point in time the total distance is calculated for each candidate phone and each syllable component.

Finally, the sequence with the smallest distance is chosen as the best match. Working with a limited data set at the moment, the average accuracy rate for the utterances we have considered is 73.9%, which is very promising.

5. FUTURE WORK

The recognition work is being continued with the immediate focus on such aspects as use of more data and speakers, and the
formalization of the evaluation procedure. In the longer term other factors will also be considered. For example, although the results of the early work are promising, if hidden Markov modeling is used to optimize the results of the recovered syllable patterns in the second stage of recognition, we expect the final recognition results will be improved. Further work is also needed to ensure the recovered syllable contours are linguistically and physiologically plausible. It remains to be seen whether or not phonotactic constraints, or patterns based on sonority contours, will also be required to assist with the phonetic labeling.

6. 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 smoothed 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 statistical models of phone sequences, or to models of the syllable as a sequence of phones. This suggests that the syllable is a good articulatory unit for speech recognition processing. Ultimately, phonemic labelling and morphological recognition must underpin the recognition process, and we consider this will be supported by syllable identification.

7. REFERENCES