Pseudo-Articulatory Representations and the Use of Syllable Structure for Speech Recognition

LI ZHANG & WILLIAM EDMONDSON
School of Computer Science
University of Birmingham
Birmingham
United Kingdom

Abstract: - The alternative approach for speech recognition proposed here is based on pseudo-articulatory representations (PARs), which can be described as approximation of distinctive features, and aims to establish a mapping between them and their acoustic specifications. This mapping which is used as the basis for recognition is first done for vowels. It is obtained using multiple regression analysis after all the vowels have been described in terms of phonetic features and an average cepstral vector has been calculated for each of them. Based on this vowel model, the PARs values are calculated for consonants. At this point recognition is performed using a brute search mechanism to derive PAR trajectories. Subsequently we'll show how a model of syllable articulation can be used with PAR trajectories to computationally provide a general articulatory transcription of speech without phonetic labeling. This will form the basis of a speech recognition system. Finally the recovered syllable patterns are used to obtain a phone sequence. The results are very promising taking into account the preliminary nature of the work and the novelty of the approach.

Key-Words: - Pseudo-articulatory representations (PARs), Hidden Markov models (HMMs), Resynthesis, Phonotactic, Sonority

1 Introduction

For the past two decades the prevailing approach to speech technology has been that of hidden Markov models (HMMs). It made it possible to improve the recognition results significantly which justified its use. However, in search of new ways of overcoming the limitations posed by HMMs, attention has been diverted more and more frequently towards exploitation of the phonetic and linguistic knowledge.

1.1 Use of distinctive features in combination with HMMs

Phonetic features are one of the most common manifestations of this knowledge and have been used by several people in combination with HMMs to optimize the recognition results and provide a more phonetically-justified approach to speech recognition. Espy-Wilson, for instance, extracts distinctive features of manner-of-articulation based on their acoustic correlates and then trains HMMs using those correlates in order to recognize semivowels [1]. Deng and Erler, on the other hand, employ phonetic features as the basic modeling unit which they use to train HMMs (a different model for each feature) and allow for asynchronous time alignment over adjacent phones [2]. Johnson models speech recognition as the estimation of distinctive feature values at articulatory landmarks and claims their superiority to phonemes [3]. Kirchoff, too, uses phonetic features to define syllable-length units which then serve as triphone models for HMM training [4].

1.2 Pseudo-articulatory representations

The research presented here attempts to show that it is possible to do away with hidden Markov modeling altogether. The approach we have taken is to develop a computational model for processing speech in a non-segmental way by using pseudo-articulatory representations which represent linguistic generalizations and idealizations of articulation and the articulator positions.

PARs are derived from linguistic specifications of articulatory activity, which are both abstract and idealized. The abstractions and idealizations permit the linguistic generality to be distinguished from the articulatory reality; this is what we need in speech processing. PARs attempt to retain the linguistic generality while also gaining some realism through adoption of continuous articulatory feature values; the latter permits mapping to acoustic values [5]. PARs, in the general case, are mappings between properties of the speech signal and parameters with physiological and/or linguistic plausibility. Their value lies in the fact that constraints on values taken by a PAR can be motivated by physiological or linguistic factors. In reality, of course, PARs are mappings between articulatory or linguistic parameters and parameters used to generate speech (eg. Klatt parameters), or which are derived from speech. The constraints provided via this mapping ensure that synthesis is sensibly controlled and that recognition yields plausible values. But there is more to be gleaned from PARs.

PARs can be described as the phonetician's idealizations of the articulatory process and are approximated by distinctive features in phonetics. Their values are, however, continuous rather than binary and range from 0 to 100. It has been demonstrated [6] that in a simple case, and using PARs mapping formants to modified distinctive features taken from phonology, it is possible to overcome the *ventriloquist effect*, where acoustic evidence from many different articulatory configurations is recognized as a single phone. In general, PARs are abstract enough to discard the acoustic intricacies of the speech signal and the irrelevant fine details of articulation, and this makes them suitable for the work on recognition.

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 [7] note that discussion of sonority as an organizing principle for syllable structure goes back to the late 19th century. More recently Kaye [8] 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'.

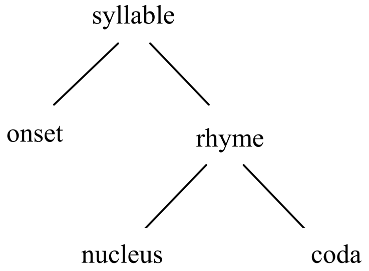

Figure 1

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 [7], 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-tar' 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).

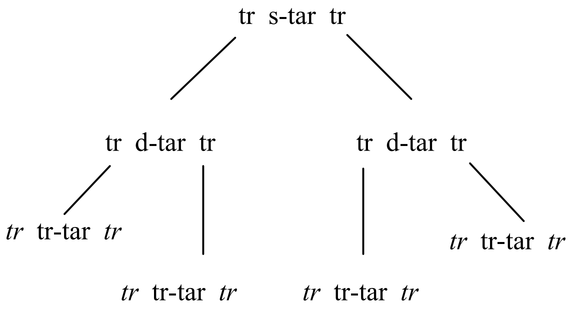

Figure 2

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-tar**, *tr*, d-tar, *tr*, **tr-tar**, *tr*
[æ] [>p] [pt] [t<]

Figure 3

This shows that the articulatory detail can be labeled '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, labeled 'd-tar' above.

3 Mapping Procedure

First of all a mapping has to be established between PARs and acoustic parameters.

Cepstral coefficients are chosen as acoustic parameters capable of describing all sound classes as opposed to previously used formant frequencies. The speech data are obtained from the TIMIT database and for the time being only one speaker is taken into account. The phone labeling is used to identify phone boundaries and for each phone a single, average vector of 18 cepstral coefficients is calculated based on all the available occurrences of this phone.

3.1 Vowel model

The mapping is done for vowels to start with. 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]. Subsequently, the vectors as well as the PAR values are used as input to multiple regression analysis in order to establish the mapping. In this way a vowel model is obtained.

3.2 PAR derivation for consonants

In order to determine PAR values for consonants an assumption is made that the production of consonants is similar to that of vowels and that they can be described using the same four features. Again an average vector of 18 cepstral coefficients is calculated for each consonant; however, this time the PAR values are not taken from phonetic textbooks, but calculated using the vowel model. A set of 18 linear equations are formed for each consonant where on the one side, there are the cepstral coefficients (cc1 to cc18) and on the other side - the a_i regression constants taken from the vowel model.

$$cc_i = a_0 + a_1h + a_2b + a_3r + a_4t + a_5hb + a_6hr + a_7ht + a_8br + a_9bt + a_{10}rt$$

A brute search mechanism is employed to find the unknown feature values in a solution space, which is gradually restricted. As a result of it, a set of four values for high, back, round and tense are determined for each consonant. At that point the mapping is complete and everything is ready to run recognition experiments.

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 is done with a fixed window sliding along the speech pattern. This output establishes every 10 msec a set of 18 cepstral coefficients for the incoming speech. Again a brute search mechanism is used (the same as in deriving PARs for consonants) which by gradually reducing the solution space determines four PAR values for each set of 18 cepstral coefficients. As a result of this, an utterance is described with 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 four pseudo-articulatory values are produced for each 10ms of speech in the test file. 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 to every segment in the transcription files. The values for vowels are taken from the vowel model, i.e. the definitions obtained from textbooks and used as data points in the mapping procedure. 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 [10] 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 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 labeled 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 this is shown diagrammatically in figure 4. The sentence is: There is usually a valve.

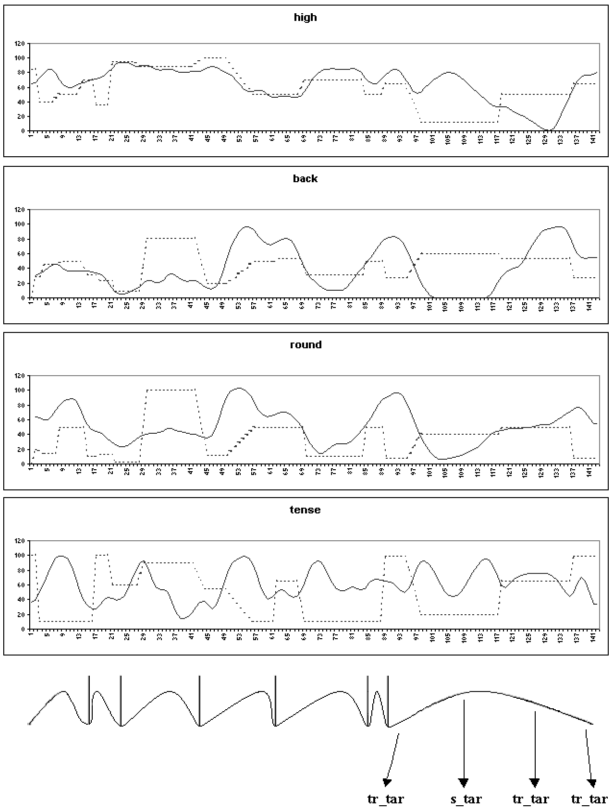

Figure 4 The top 8 traces respectively show the idealized feature trajectories (dotted line) and smoothed feature trajectories (continuous line) of high, back, round, tense. The bottom section shows in schematic form the recovered syllable positions from the smoothed trajectories and the syllable structure for one example.

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 unlabeled 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 labeling and morphological recognition must underpin the recognition process, and we consider this will be supported by syllable identification.

Finally, using PARs offers a higher level of abstraction than statistical approaches and thus a good chance of successfully dealing with the problem of many-to-one mappings. Since PARs are allowed to overlap and take continuous values, there is no need for rigorous segmentation. That should allow us to solve the problem of coarticulation. Finally, this approach is fundamentally inherent within the process of speech articulation and reflects directly the current state of phonetic knowledge.

References:

[1] Espy-Wilson, C. Y. A Feature-Based Semivowel Recognition System. J. Acoust. Soc. Am., Vol. 96, 1994.

[2] Deng, L. and Erler, K. Structured Design of a Hidden Markov Model Speech Recognizer Using Multivalued Phonetic Features. J. Acoust. Soc. Am., Vol. 92, 1992.

[3] Johnson, M. E. Automatic Context-Sensitive Measurement of the Acoustic Correlates of Distinctive Features at Landmarks. Proceedings of ICSLP'94, 3:1663-1642, 1994.

[4] Kirchoff, K. Syllable-Level Desynchronisation of Phonetic Features for Speech Recognition. Proceedings of ICSLP'96, 4:2274-2276, 1996.

[5] Edmondson, W.H., Iskra, D. J. and Kienzle, P. Pseudo-Articulatory Representations: Promise, Progress and Problems. Proceedings of EUROSPEECH'99, 3:1435-1438, 1999.

[6] Iles, J.P., & Edmondson, W.H. Quasi-Articulatory Formant Synthesis. Proceedings of ICSLP'94, 3:1663-1666, 1994.

[7] Bell, A. & Hooper, J. B. Issues and Evidence in Syllabic Phonology. In Syllables and Segments, A. Bell and J. B. Hooper (eds.), pp. 1-22. Amsterdam: North Holland, 1980.

[8] Kaye, J. Phonology: A Cognitive View. New Jersey: Lawrence Earlbaum Associates, 1989.

[9] Ladefoged, P. A Course in Phonetics. Harcourt Brace Jovanovich, 1975.

[10] Edmondson, W. H., & Zhang, L. Pseudo-Articulatory Representations and the Recognition of Syllable Patterns in Speech. Proceedings of EUROSPEECH'01, 1:595-598, 2001.