Feature-Based Approach to 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 that 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 vowel and consonant models to derive idealized PAR trajectories. And we'll show how a model of syllable articulation can be used with PARs to computationally provide a general articulatory transcription of speech without phonetic labeling. This will form the basis of a speech recognition system.

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

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 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.

2.1 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 [7]. 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.

2.2 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.

$$cc_i = a_0 + a_1 h + a_2 b + a_3 r + a_4 t + a_5 hb + a_6 hr + a_7 ht + a_8 br + a_9 bt + a_{10} rt$$

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.

3 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 [8] note that discussion of sonority as an organizing principle for syllable structure goes back to the late 19th century. More recently Kaye [9] 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.

3.1 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 1, 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 2 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, 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*

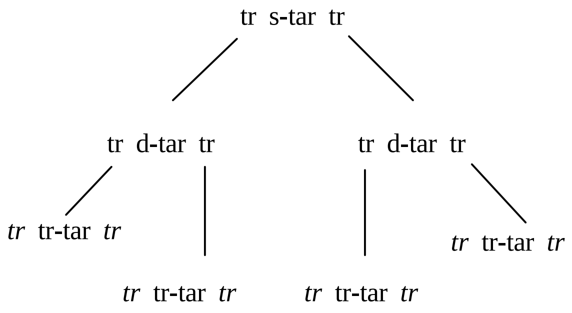

Figure 1

An example of how this might be used for the English word ‘apt’, is shown in Figure 2.

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

Figure 2

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.

4 Use of PARs

We now show how the details of syllable articulation can be recovered from pseudo-articulatory representations.

We choose 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 demonstrate the feasibility of the relationship, we will go on to consider the problems of computationally derived PARs.

The idealized PARs are produced by ascribing four feature values to every segment in the transcription files. The values for vowels are taken from the vowel model. The values for consonants are taken from the consonant model, which we have discussed in 2. 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 3.1, 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 are shown diagrammatically in figure 3 and in detail in figure 4.

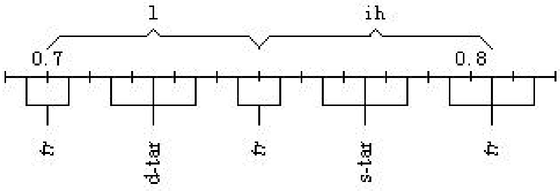

Figure 4 the analysis results between 0.7s and 0.8s by every 10ms

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

5 Results

The results are evaluated at different points in the recognition process. As a result of the regression analysis, not only are the regression constants obtained, but the coefficients of determination as well. These coefficients are nearly 1 for all the cepstral coefficients implying that there is very little difference between the estimated and the actual values. That means also that the equation obtained in this way fits the data very well.

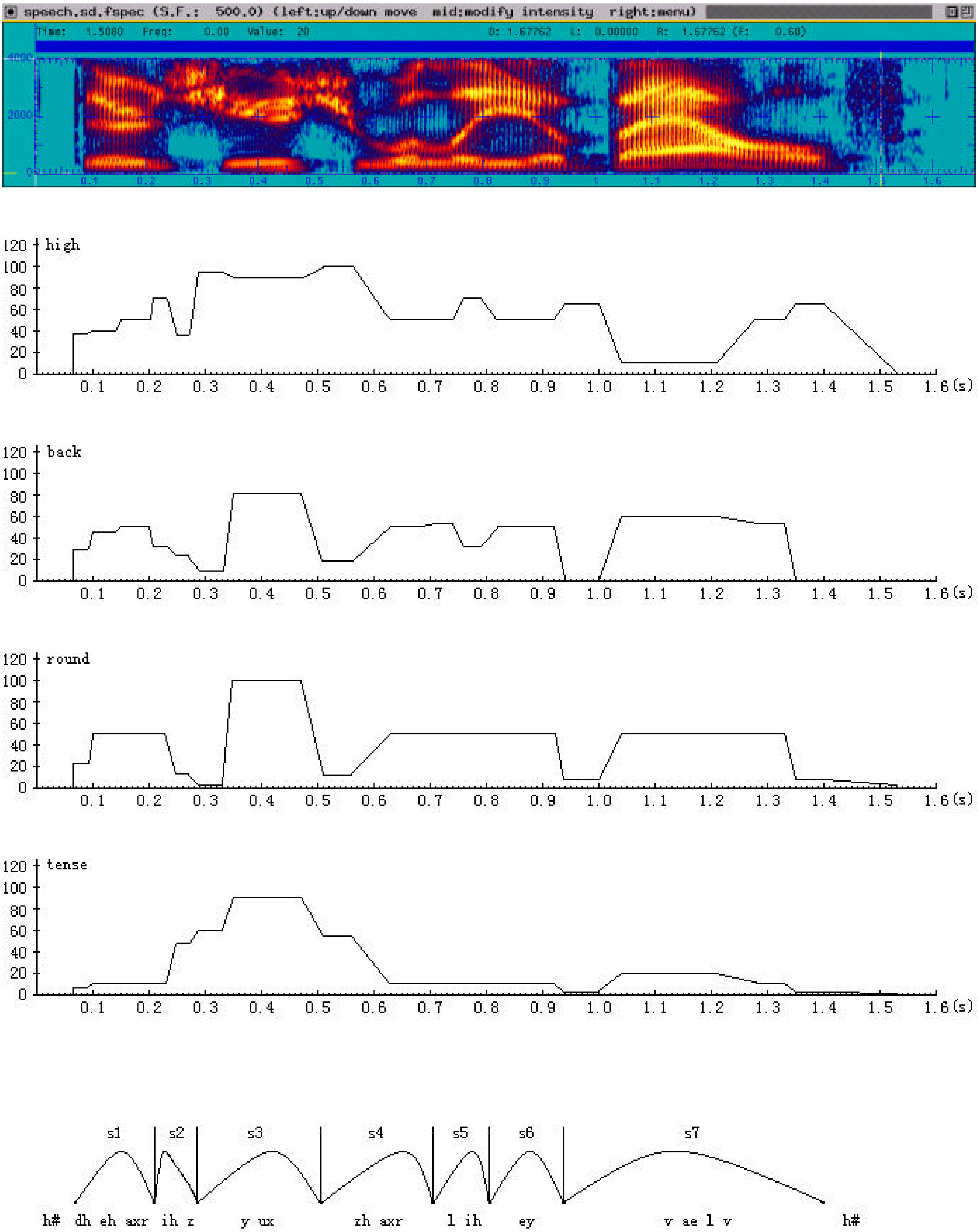

Figure 3

The top section shows the spectrogram of the utterance " There is usually a valve." .
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.

5.1 Evaluation of the mapping procedure

In order to evaluate the mapping procedure, the PAR values obtained for consonants are compared to phonetic feature specifications found in textbooks [10]. The feature values given in books are always binary, so in order to make the comparison possible [-] is assumed to correspond to all the values in the range 0-33, [-+] to the range 34-66, and [+] to 67-100. If a found PAR value falls within this range, it is considered to be 'the right match'. The number of right matches is highest for the feature 'round' (20 out of the total of 29 consonants taken into account in the analysis), followed by 'high' and 'back' (both 14), and lowest for 'tense' (9). These results may seem not too promising, but a closer observation makes it clear that some of the PAR values fall just outside the given range. They are not regarded as 'the right matches', but in reality they are very close. The feature 'tense' scores lowest implying that it is the hardest one to predict from the cepstral parameters.

6 Future Work

The next step in our work will be to attempt syllable recovery using computationally derived PARs, in the manner of Iles [6], 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 labeling of the syllables. Ultimately, phonemic labeling and morphological recognition must underpin the recognition process, and we consider this will be supported by syllable identification.

In addition, the recognition work is being continued with the focus on such aspects as optimization of the experimental setup, use of more data and speakers, and the formalization of the evaluation procedure.

7 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 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 suggests that the syllable is the right articulatory unit for speech recognition processing.

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. and Edmondson, W.H. Quasi-Articulatory Formant Synthesis. Proceedings of ICSLP'94, 3:1663-1666, 1994.

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

[8] Bell, A. and 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.

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

[10] Atkinson, M., Kilby, D. and Rocca, I. Foundations of General Linguistics, Unwin Hyman, London, 1991.