Proceedings of the Institute of Acoustics

FEATURE-BASED APPROACH TO SPEECH RECOGNITION

D J Iskra and W H Edmondson

University of Birmingham, School of Computer Science, Birmingham, UK

E-mail: d.j.iskra@cs.bham.ac.uk, w.h.edmondson@cs.bham.ac.uk

1 ABSTRACT

Pseudo-articulatory representations (PARs) can be described as approximations of distinctive features. The alternative approach to speech recognition proposed here is based on PARs and aims to establish a mapping between them and their acoustic specifications (in this case cepstral coefficients). This mapping which serves as the basis for recognition is first obtained for vowels 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. Next, using the regression coefficients and the respective average cepstral vectors, the PAR values are calculated for consonants. Now that the mapping is complete, recognition is performed on speech data for a single speaker using a brute search mechanism to derive PAR trajectories and subsequently dynamic programming to obtain a phone sequence. The results are not as good as when hidden Markov modelling is applied, but they are very promising if we take into account the early stage of the experiments and the novelty of the approach.

2 BACKGROUND

For the past two decades the prevailing approach to speech recognition has been that of hidden Markov models (HMMs). This very powerful statistical technique made use of the constantly improving computing resources and allowed for speaker-independent continuous speech recognition with the results unheard of before. However, even hidden Markov modelling knows its limitations, mostly inherent in its purely statistical nature, the fact which has recently diverted attention more and more frequently back towards exploitation of the phonetic and linguistic knowledge.

2.1 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 optimise 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 recognise semivowels [1]. Deng and Erler, on the other hand, employ phonetic features as the basic modelling 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].

2.2 Pseudo-Articulatory Representations

The research presented here attempts to show that it is possible to do away with hidden Markov modelling altogether. The approach is based on pseudo-articulatory representations - the idea which was introduced some time ago by Iles and Edmondson [5]. PARs can be defined as the phonetician's idealisations of the articulatory process and are approximated by distinctive features in phonetic. Their values are, however, continuous rather than binary and range from 0 to 100. Iles applied this idea initially to synthesis [6]. With the aim of driving speech synthesis in an articulatory manner he established a mapping between PARs and acoustic specifications such as formant frequencies, bandwidths and amplitudes for cardinal vowels. Then he used PARs as input to control a formant-based (Klatt) synthesiser. Encouraged by the synthesis results, he tried inverse mapping to determine articulatory trajectories from the incoming signal. This was done using brute search on the previously established formant trajectories. The phone inventory was restricted to vowels and semivowels, but the results were very promising and the resynthesised PARs provided a close match to the original [7]. This idea is being continued further.

Because of their abstractness PARs make it possible to discard the acoustic intricacies of the speech signal and the irrelevant fine details of articulation. This is also what makes them as suitable for work on recognition as on synthesis.

3 MAPPING PROCEDURE

The core of the approach lies in the mapping of PARs onto suitable acoustic parameters and this had to be established first of all. The limitations of formant frequencies led us to seek an alternative way of acoustic representation of speech and subsequently choose cepstral coefficients as capable of describing all classes of sounds. The speech data were obtained from the TIMIT database and at that stage only one speaker was taken into account. The TIMIT label files were used to determine phone boundaries and for each phone a single, average vector of 18 cepstral coefficients was calculated based on all the available occurrences of this particular phone. The cepstral analysis was performed using Entropic's ESPS software.

3.1 Vowel Model

The mapping was done for vowels to start with. Apart from cepstral coefficients a PAR description was also required and that was obtained by selecting four features: high, back, round, tense and ascribing for each of them a value between 0 and 100 to every vowel based on the measurements provided by Ladefoged [8]. Subsequently, the cepstral as well as the PAR vectors were used as input to multiple regression analysis in order to establish the mapping. The multiple regression analysis was based on a set of 18 linear coefficients

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

where cc_i were the successive cepstral coefficients, *h, b, r, t* - the values for high, back, round and tense, and the a_i – the regression coefficients. Therefore, the values of the regression coefficients, which were calculated in the course of this analysis, provided the link between cepstrai coefficients and PARs. In this way a vowel model was obtained.

3.2 PAR Derivation for Consonants

In order to determine PAR values for consonants an assumption was 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 was calculated for each consonant; however, this time the PAR values were not taken from phonetic textbooks, but calculated using the vowel model. The same set of linear equations was used as well as the a_i regression coefficients from the vowel model, except that now the values for *h, b, r, t* were the unknown. To find those a brute search mechanism was employed which gradually restricted the solution space until it arrived at four PAR values for each consonant. At that point the mapping was complete and everything needed to run recognition experiments was in its place.

4 RECOGNITION

In the recognition process two successive stages could be clearly distinguished. In the first stage the transition from the acoustic representation of the incoming signal to the pseudo-articulatory one was made, in other words, feature trajectories were determined on the basis of cepstral vectors. The second stage comprised the movement from the pseudo-articulatory to the phone level of description and, using the PAR values as input, resulted in a sequence of phone labels (see figure 1).

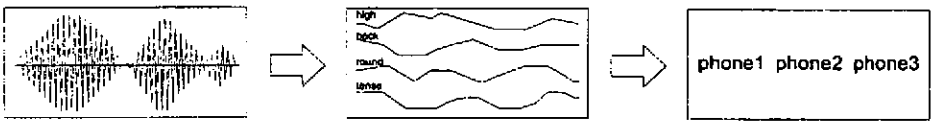

Figure 1. Different manners of representation of the speech signal during the recognition process.

4.1 Transition from the Acoustic to the Pseudo-Articulatory Level

The first stage of the recognition was done with a fixed window sliding along the speech pattern. This output established every 10 ms a set of 18 cepstral coefficients for the incoming speech. The cepstral coefficients were used as input in a brute search algorithm (the same as in deriving PARs for consonants) which by gradually reducing the solution space determined four PAR values for each set of 18 cepstral coefficients. As a result of this, an utterance was represented with a set of new values for high, back, round, tense every 10 ms. When plotted, these values formed feature trajectories for that utterance.

4.2 Finding a Phone Sequence

After the feature trajectories had been determined, dynamic programming was used [9] in order to find the best matching sequence of phones. The distance was calculated between each set of four incoming feature values and the reference feature values for every phone. The duration information was used to modify the distances and at each point in time the total distance was calculated for each phone and each starting point. Finally, the sequence with the smallest distance was chosen as the best match.

5 RECOGNITION RESULTS

The results were evaluated at different points in the recognition process. For example, as a result of the regression analysis, not only were the regression constants produced, but the coefficients of determination as well. These coefficients were nearly 1 for all the cepstral coefficients implying that there was very little difference between the estimated and the actual values, and that the equation obtained in this way fitted the data very well.

5.1 Evaluation of the Mapping Procedure

In order to evaluate the mapping procedure, the PAR values obtained for consonants were 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 [-] was 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 fell within this range, it was considered to be "the right match". The number of right matches was 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 lower than expected, but a closer observation made it clear that some of the PAR values fell just outside the given range. They were not regarded as "the right matches", but in reality they were very close. The feature "tense" scored lowest implying that it is the hardest one to predict from the cepstral parameters.

5.2 Resynthesis

Another way of determining how successful the mapping procedure had been was resynthesis. The PAR values obtained as a result of the recognition process were substituted into the equations together with the regression coefficients and new cepstral coefficients were calculated. These were then used as the basis for synthesis and compared to the original files by listening to both. The quality of the resynthesised utterances was judged to be very good and not significantly different from the original. Additionally, the cepstral coefficients were plotted and the respective new and original coefficients were overlaid. Except for three coefficients, both the re-calculated and the original trajectories were very similar. The first four plots are shown in figure 2.

5.3 Calculating Phone Recognition Percentage

In order to evaluate the recognition results, an approach was taken of expanding the phone labels over their duration. Therefore, if a phone had been labelled to last 60 ms (whether it was the original utterance or the recognised one), it would be counted as 6 "occurrences" of

the same phone (10 ms each). This approach was meant to evaluate not only the recognition of the phone, but to take into account its duration as well. Then a percentage was calculated by dividing the number of correctly recognised phones by the number of all occurrences of this phone in the original utterances. The numbers were very different for different phones. The vowels scored highest, and among them the long vowels with 80% recognised correctly for /aa/, 88% for /uh/. The nasals and the semivowels followed with, e.g., 44% for /ng/. Some of the stops were recognised pretty well with, e.g. /bcl/ - 68%, but the other results were lower. On the whole, the fricatives and the affricates did not do very well.

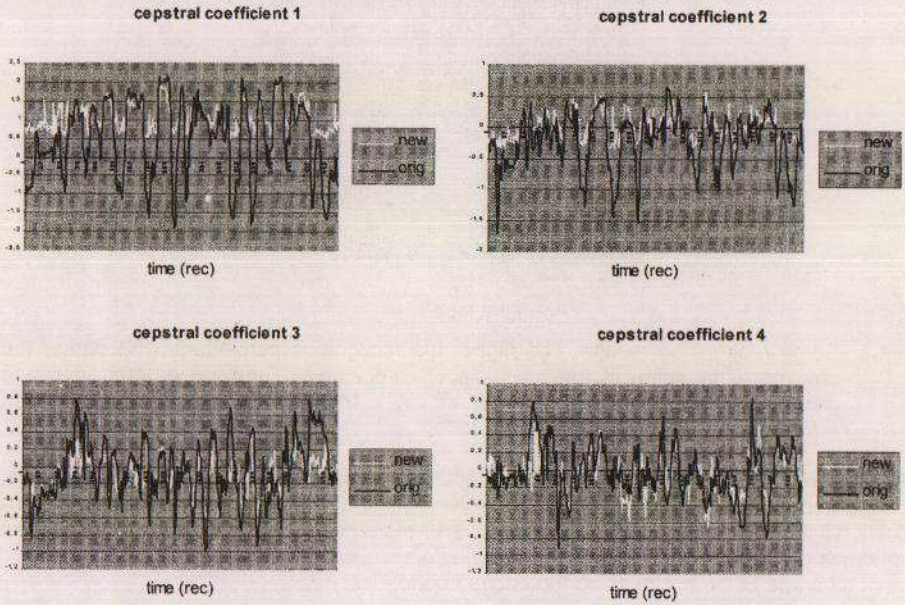

Figure 2. The original and the new cepstral trajectories for the first four coefficients and a single utterance.

6 DISCUSSION

The fact that it proved possible to perform inverse mapping and on the basis of PAR trajectories obtain new cepstral coefficients signifies that the mapping was correct. This is further enhanced by the quality of resynthesis which was very good.

As far as the recognition results are concerned, it is clear that some classes of sounds were recognised better than others. Vowels, semivowels and nasals had the best scores, which was not unexpected. These are the classes of sounds well known for their consistency, clarity

and steadiness in their phonetic realisation. These are also the sounds which can be described most adequately with the features selected earlier (high, back, etc.). Not surprisingly either, the plosive and the fricative sounds pose major problems, which is a case well known in automatic speech recognition and is due to the acoustic nature of these sounds. Therefore, future efforts to improve the recognition results will concentrate on these classes of sounds.

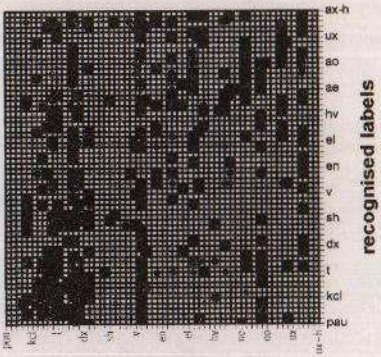

Figure 3. Some recognition results. The higher the recognition percentage, the darker the shading. Only some of the phone labels are visible. They are ordered in sound classes with silence/noise, plosives, affricates, fricatives, nasals, semivowels, and vowels from left to right and bottom to top.

The evaluation procedure used here was not optimal. The smallest chunk of labelled speech was regarded to be 10 ms. Therefore, if the duration of a phone was, e.g., 57 ms, for the evaluation it would be assumed to stretch over 6 10-ms windows, the same as the phone with the duration of 63 ms. In reality, however, this difference could be quite significant and could account for some of the mistakes on the phone boundaries.

The recognition work is being continued with the focus on such aspects as optimisation of the experimental setup, use of more data and speakers, and the formalisation of the evaluation procedure.

7 CONCLUSIONS

The initial results are lower than those obtained using hidden Markov models, but taking into account the fact that this is a completely different approach, they are still regarded as very promising at this stage of experiments.

Using PARs offers several advantages. It moves recognition to a higher level of abstraction than statistical approaches and thus makes it possible to deal successfully 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. Consequently, smooth transitions from one segment to another should allow us to solve the problem of coarticulation. Finally, this approach is fundamentally inherent within the process of speech articulation and makes direct use of phonetic knowledge.

8 ACKNOWLEDGMENTS

The authors wish to thank Prof. Martin Russell, School of Electrical Engineering, University of Birmingham for his help and valuable comments.

9 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. Iles, J. P. and Edmondson, W. H. "Control of speech synthesis using phonetic features", *Proceedings of the Institute of Acoustics Autumn Conference on Speech and Hearing*, 14:369-373, 1992.
6. Iles, J. P. *Text-to-speech Conversion Using Feature-Based Formant Synthesis in a Non-Linear Framework*, PhD thesis, School of Computer Science, University of Birmingham, 1995.
7. Iles, J. P. and Edmondson, W. H. "Quasi-articulatory formant synthesis", *Proceedings of ICSLP'94*, 3:1639-1642, 1994.
8. Ladefoged, P. *A Course in Phonetics*, Harcourt Brace Jovanovich, Inc., 1975.
9. Sakoe, H. and Chiba, S. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Trans. ASSP*, 26:43-49, 1978.
10. Atkinson, M., Kilby, D. and Rocca, I. *Foundations of General Linguistics*, Unwin Hyman, London, 1991.