FEATURE-BASED APPROACH TO SPEECH RECOGNITION

Dorota J. Iskra and William H. Edmondson

School of Computer Science, University of Birmigham
United Kingdom
Email: d.j.iskra@cs.bham.ac.uk, w.h.edmondson@cs.bham.ac.uk

ABSTRACT

The alternative approach to speech recognition proposed here is based on pseudo-articulatory representations (PARs), which can be described as approximations of distinctive features, and aims to establish a mapping between them and their acoustic specifications (in this case cepstral coefficients). 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 PAR values are calculated for consonants. At this point recognition is performed 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 used, but very promising taking into account the early stage of the experiments and the novelty of the approach.

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. Recently, 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 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].

1.2 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] and was initially applied to speech synthesis by Iles [6]. 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. In his work on synthesis Iles established a mapping between PARs and acoustic specifications (formants, bandwidths, amplitudes) for all sounds and then used PARs to drive a formant-based synthesizer in a more articulatory manner. He also attempted the inverse mapping and obtained some recognition results for vowels and semivowels [7]. This idea has been continued further. PARs are abstract enough to discard the acoustic intricacies of the speech signal and the irrelevant fine details of articulation, and this makes them equally suitable for work on recognition as well as synthesis.

2. MAPPING PROCEDURE

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

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

2.1 Vowel Model

The mapping was done for vowels to start with. The PAR description was 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 [8].

Subsequently, the vectors as well as the PAR values were used as input to multiple regression analysis in order to establish the mapping. In this way a vowel model was obtained.

2.2 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. A set of 18 linear equations were formed for each consonant where on the one side, there were the cepstral coefficients (cc_1 to cc_{18}) 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 was employed to find the unknown feature values in a solution space which was gradually restricted. As a result of it, a set of four values for high, back, round and tense were determined for each consonant. At that point the mapping was complete and everything was ready to run recognition experiments.

3. RECOGNITION

In the recognition process two successive stages could be clearly distinguished. The first stage was responsible for the transition from the acoustic representation of the incoming signal to the pseudo-articulatory one with feature trajectories as a result of this stage of recognition. The second stage concerned the movement from the pseudo-articulatory to the phone level of description and produced a sequence of phone labels.

3.1 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 msec a set of 18 cepstral coefficients for the incoming speech. Again a brute search mechanism was used (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 described with a set of values for high, back, round, tense every 10 msec. When plotted, these values presented feature trajectories for that utterance.

3.2 Finding a Phone Sequence

At that point dynamic programming was used [9] in order to find the best matching sequence of phones by calculating the distance between each set of four incoming feature values and the reference table. 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.

4. RESULTS

The results were evaluated at different points in the recognition process. As a result of the regression analysis, not only were the regression constants obtained, 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. That meant also that the equation obtained in this way fitted the data very well.

4.1 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 not too promising, 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.

4.2 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 was labelled to last 60 msec (whether it was the original utterance or the recognised one), it would be counted as 6 “occurrences” of the same phone (10 msec each). This 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 recognized 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% recognized correctly for /aa/, 88% for /uw/. The nasals and the semivowels followed with, e.g., 44% for /ng/. Some of the stops were recognized 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.

It is clear that some classes of sounds were recognized better than others, which was not unexpected. Therefore, vowels, semivowels and nasals had the best scores. These are the classes of sounds well-known for their consistency, clarity and steadiness in their phonetic realization. These are also the sounds which can be described most adequately with the features selected earlier (high, back, etc.). Not surprisingly, 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.

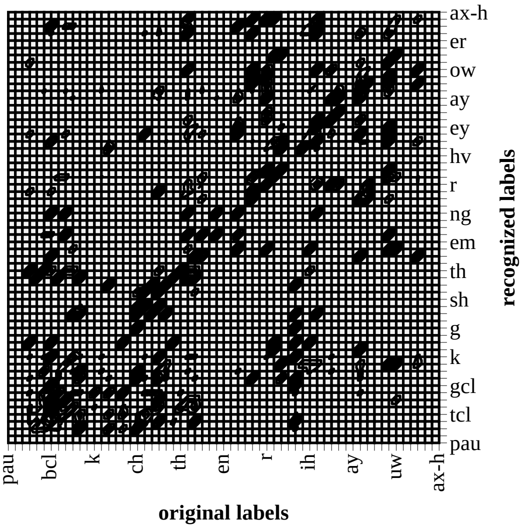

Figure 1: Some recognition results. The higher the recognition percentage, the darker the shading. Only some of the phone labels are visible. They are ordered in sound classes with silence/noise, plosives, affricates, fricatives, nasals, semivowels, and vowels from left to right and bottom to top.

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

5. FUTURE WORK

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. 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.

6. CONCLUSIONS

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.

7. ACKNOWLEDGMENTS

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

8. 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.