USING ARTICULATORY KNOWLEDGE IN AUTOMATIC SPEECH RECOGNITION

Helmer Strik
Dept. of Language & Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
E-mail : STRIK@LET.KUN.NL

1. introduction

Over the years different types of speech recognizers have been proposed and tested. During the last decade (or maybe even longer) hidden Markov models (HMMs) seem to have a better performance than other types of speech recognizers, like e.g. rule-based speech recognizers. This state of affairs has led to a gap between speech technology on the one hand, and phonetics and phonology on the other. Clearly, this is not an ideal situation, because both fields could and should benefit from each other. The important question then is: why is the performance of rule-based recognizers not as good as that of HMM-based recognizers? Probably this is due to a combination of several factors like: (1) at the moment, there isn't enough knowledge available; (2) most of the knowledge available is derived from lab speech and cannot always be generalized to other types of speech; (3) the knowledge available is generally transformed into deterministic rules in rule-based systems, while HMM-based systems mainly use a kind of stochastic rules; (4) and finally, and probably most important, in many rule-based recognizers local decisions often have to be made (i.e. whether a segment is voiced or not), while in HMM-based systems one overall probabilistic decision is made.

Apart from a good performance, HMMs have another important advantage, namely that training and testing of these recognizers can be done almost completely automatically. However, they also have some disadvantages, some of which are mentioned here. First of all, it is difficult to incorporate knowledge in an HMM. In certain cases it is known that the way in which HMMs model speech is not correct (a good example are the time distributions), but it often is difficult to use this knowledge in HMMs. Another disadvantage of HMMs is the enormous amount of training material needed to build the models. Furthermore, in practice a new database is usually collected whenever a new application is needed. The reason that such large databases are needed is that HMMs contain many parameters which have to be trained. Usually, doubling the training material gives a substantial improvement, probably because the HMMs are undertrained. However, recently Kubala et al. (1994) showed that doubling the training material only leads to a marginal improvement in the performance of their recognizer. Therefore, it could well be that conventional HMMs are reaching their maximum level of performance. At the moment their performance is good, but for some domains it is not good enough yet (e.g. for informal speech, as in the Switchboard corpus).

Therefore, and for other reasons expressed below, we decided to test a new approach which is briefly described in section 2. In section 3 the advantages and disadvantages of the HMMs used in the new method (which will be called simply new HMMs in this article) are touched on. I will finish the present section by mentioning the goals of our research.

Our most important goal is (1) to bridge the gap between speech technology and phonetics-phonology, which was mentioned above. We try to do this by using a more realistic model of speech production, as will be described in the next section. In this way we hope to achieve two other goals, viz. (2) to obtain (statistical) knowledge from large amounts of 'natural speech' (as opposed to 'lab speech', on which most knowledge is based now); and (3) to improve speech recognition.

The goals are deliberately presented in this order, because we think that by using a more realistic model and by incorporating knowledge in this model recognition performance will increase in the long run. However, at the beginning there could be a decrease in performance, because since there is less experience with these new HMMs, they will not be as finely tuned as conventional HMMs are.

To conclude this section, I would like to point out that the research presented in this article will mainly

be conducted by the author in the framework of a post-doctoral KNAW-project. Since this research is in a very early stage, all comments are welcome.

2. the new method

In this section I would like to give a brief description of the method we intend to use. This new approach is based on the work of Deng and colleagues. A more thorough description can be found in Deng & Erler (1992) and Deng & Sun (1994). The method has been tested for (American) English. For that language it gave good results. We intend to test this method for Dutch.

The difference between the conventional and the new HMMs is depicted in Figure 1. In conventional HMMs a grapheme string is converted into a phoneme string (or more generally, a string of speech units). Based on this phoneme string, a state-transition network is constructed. These are the usual steps in top-down processing. From the bottom-up side the speech signal is coded into a set of speech parameters. Finally, the relation between the networks and the speech parameters is modelled stochastically. In the new approach an extra layer is inserted in the top-down side. The phoneme string is converted into a feature-overlap pattern, which, in turn, is transformed into a network. How this is done is explained below. The explanation is divided in six steps, which are the stages that have to be passed through when using this method for a given lexicon.

1. The first thing to do is to define or select a set of articulatory features. It is important to have a feature set which can describe both consonants and vowels. Some of these feature sets have been proposed in the literature, (see e.g. Browman & Goldstein, 1992; Clements, 1993; Deng & Sun, 1994). Although we intend to compare different feature sets at a later stage, we will start by using the same set as Deng & Sun (1994), which is motivated by the work of Browman & Goldstein (1992). This set consists of five multi-valued features: lips, tongue tip, tongue body, velum and glottis.

2. Each word in the lexicon has to be described in terms of 'stationary' speech units. This can be done partly automatically by using available lexica that contain grapheme and phoneme information, and by using a grapheme-to-phoneme conversion tool. For (American) English Deng and colleagues based their choice on the 61 (quasi-)phonemic labels of the TIMIT database. For Dutch we will have to find out what the optimal set of speech units is. As a first try we will use the labels from the Dutch Polyphone database.

3. A list has to be drawn up which contains the values of the features for all speech units. The values in this list are the values of the features for context-independent speech units, i.e. they can be thought of as the target values for the speech units. A value of zero means that the feature is unmarked, i.e. that in producing a speech unit this specific feature is not important. A list of this kind has to be made once for each feature set. In principal it is manual work, but we will try to make it semi-automatic.

4. For each speech unit in context a feature-overlap pattern has to be constructed and subsequently transformed into a network.

??? korte uitleg over feature-overlap pattern

5. The next step is the construction of networks for complete utterances. These networks are simply the concatenation of the networks for the individual context-dependent speech units, as is the case with conventional HMMs.

6. Finally, the models have to be trained and tested. As in conventional HMMs, this is done in a Bayesian framework. In fact, the algorithms for training and testing are only slightly different from those used for conventional HMMs (see Deng & Erler, 1992).

3. discussion

In the introduction I already mentioned some of the advantages and disadvantages of conventional HMMs. The pros and cons of the method proposed in this article are presented here. A disadvantage of the new HMMs is that training is not completely automatic. In the previous section I already mentioned that some of the work is (partly) manual: the feature values have to be defined for all speech units (once), and part of the database has to be labelled for bootstrapping of the training procedure (also once). Although part of this work will be manual, we will try to make a large part of it (semi-)automatic.

The new HMMs also have a number of important advantages. First of all, with the new HMMs data can be shared between states of different speech units. The amount of data sharing that can be achieved in this way is generally larger than with conventional HMMs, and it is based on articulation. Another advantage is that context dependencies are modelled explicitly in an articulatory framework, and that knowledge about e.g. co-articulation can be incorporated directly in the model. In general, knowledge can be used to define rules which can constrain the construction of the feature-overlap pattern. Although knowledge is used in constructing the model, a substantial part of the model remains stochastic. Also in these HMMs, as in the conventional HMMs, one global probabilistic decision is made. This is a clear advantage compared to other rule-based recognizers in which wrong local decisions often lead to catastrophic errors. In fact, these new HMMs should be situated somewhat between the conventional rule-based systems that are almost completely deterministic, and the conventional HMMs which are mainly stochastic. Finally, these new HMMs can be used to obtain (statistical) knowledge for large amounts of speech. This can be done in the following manner. After training with a large amount of utterances a Viterbi decoding can be used to find the optimal sequence of states for each utterance. From these optimal sequences one can infer in which cases and how often articulatory features spread, the amount of overlap, the timing between the different articulatory features, and also other things.

To sum up, the new HMMs make it possible to combine the statistical insights gained in research with conventional HMMs, with the knowledge accumulated in so many years of research in phonetics and phonology. We are convinced that this approach is worth trying and expect that in the long run it will lead to better performance in speech recognition.

Acknowledgements

The research of Dr. W.A.J. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

References

Browman, C. & Goldstein, L. (1992) Articulatory phonology: An overview. Phonetica 49, pp. 155-180.

Clements, G.N. (1993) Lieu d'articulation des consonnes et des voyelles: une theori'e unifi'ee. In: Bernard Laks & Annie Rialland (eds.) Architecture des Representations Phonologiques. Paris, CNRS., pp. 101-145.

Deng, L. & Erler, K. (1992) Structural design of a hidden Markov model based speech recognizer using multivalued phonetic features: Comparison with segmental speech units. J. Ac. Soc. Am. 92 (6), pp. 3058-3067.

Deng, L. & Sun, D.X. (1994) A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J. Ac. Soc. Am. 95 (5), pp. 2702-2719.

Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R. & Zavaliagkos, G. (1994) Comparative experiments on large vocabulary recognition. Proc. of the IEEE ICASSP, Adelaide, South Australia, Vol. I, pp. 561-564.