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

When people produce repetitions of the same utterance, the resulting speech signals belonging to these repetitions are usually very different. This is a well-known fact. Generally, the large amount of variation in the speech signals is not problematic for human listeners, but it certainly is for automatic speech recognizers.

Over the years researchers have proposed various solutions to the problems of automatic speech recognition. As a result, different types of speech recognizers have been developed. During the last decade it has appeared that speech recognizers based on hidden Markov models (HMMs) usually have a better performance than other types of speech recognizers, like e.g. rule-based speech recognizers. Probably this is due to a combination of several factors. Two important differences between rule-based and HMM-based recognizers are (1) that the first use mainly deterministic rules while the latter use a kind of stochastic rules; and (2) that in rule-based recognizers local decisions are often 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 advantage, namely that training and testing of these recognizers can be done almost completely automatically. This is certainly a big advantage because large amounts of data are needed to train the numerous parameters in the HMMs. In general, doubling the training material gives a substantial improvement, which seems to suggest that the HMMs are undertrained. However, recently Kubala et al. (1994) have shown that increasing the training material from 12 to 24 hours of speech only leads to a marginal improvement in the performance of their recognizer. They concluded that the HMMs (they used) were not able to take full advantage of the additional training data. Therefore, it could well be that conventional HMMs are now reaching their maximum level of performance.

HMMs are models in which phonetic/phonological knowledge usually plays a limited role, at least in the conventional type of HMMs. The fact that during the last decade research on speech recognition has been confined almost completely to conventional HMMs, has had the effect of widening the 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.

Based on the considerations mentioned above, we decided to test a new approach to speech recognition. Also in this new method one global probabilistic decision is made, since this has proven to be successful (as noted above). However, the HMMs used in our approach (which will be called simply new HMMs here) are essentially different from conventional HMMs. A brief description of this new method is given in the following section.

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 mentioned above. We try to do this by using a model (i.e. the new HMM) which probably represents speech production in a more realistic way than the conventional HMM, as will be described in the next section. In this way we hope to achieve two other goals, viz. (2) to obtain (statistical) knowledge about articulation 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 articulatory 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.

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. The English examples used below are taken from Deng & Sun (1994). 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.
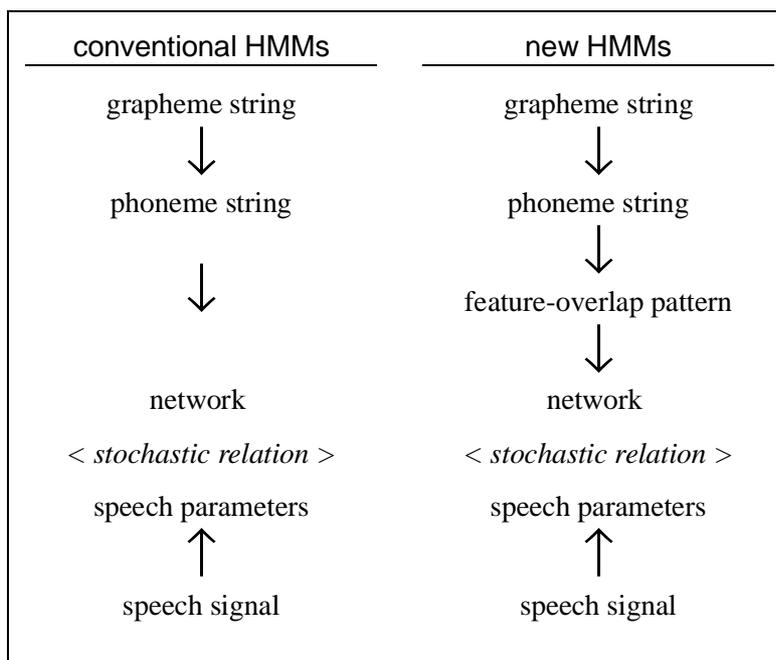
conventional HMMs	new HMMs
grapheme string	grapheme string
↓	↓
phoneme string	phoneme string
	↓
	feature-overlap pattern
↓	↓
network	network
< *stochastic relation* >	< *stochastic relation* >
speech parameters	speech parameters
↑	↑
speech signal	speech signal

Figure 1. The general structure of a conventional and a new HMM.

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.

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 (in other words, the same speech unit can be produced with different values of that feature, thus giving rise to different phonetic realisations). A list of this kind has to be made once for each feature set.

4. For each speech unit in context a feature-overlap pattern has to be constructed and subsequently transformed into a network. In Figure 2 a feature-overlap pattern is shown for the English word /ten/. Indicated are the boundaries of the three segments and the values of the features. It can be observed that some of the feature values spread into their neighbours. For instance, the tongue tip value of 1 spreads from the /t/ into the /e/, and the same value for /n/ also spreads into the /e/.

Next, the feature-overlap pattern has to be transformed to a network. In Figure 3 an example is given of a network for the /e/ in /ten/. In state A of this network the value of the feature tongue tip is L(1), which means that for this feature a value of 1 spreads from its Left neighbour (in this case the /t/). The L(2) value for glottis indicates that for the feature glottis a value of 2 spreads from the Left neighbour. Analogously, the value R(2) for velum in state G is caused by the spreading of value 2 from the Right context (the /n/) into the /e/. The result of this overlap would be that the last part of the /e/ is nasalized. In Figure 2 the spreading from /t/ into /e/ of the tongue tip value of 1 and glottis value of 2 end simultaneously. This was done for the sake of simplicity. If the amount of the two overlaps were always the same, then states B and C would not be necessary. However, in practice this is not the case. The spreading of the tongue tip feature could end after the spreading of the glottis, as is modelled by state B. State C is used to model the case in which the overlap of the tongue tip is smaller than that of the glottis.

Rules could be used to control the overlap of the features. Reliable knowledge can be used to derive these rules. For instance, if it is known that in a /t/-/e/ sequence the spreading of the tongue tip *always* ends before the spreading of the glottis, state B could be discarded. Or, alternatively, if it is known that in the same sequence none of the feature values ever spreads, state A, B and C could all be omitted. One very general rule is that the feature value zero (meaning unmarked) can never spread into a neighbour, while on the other hand it is very likely that values of the context overlap the value zero (as depicted in Figure 2).

Figure 2. A feature-overlap pattern for the utterance /ten/.
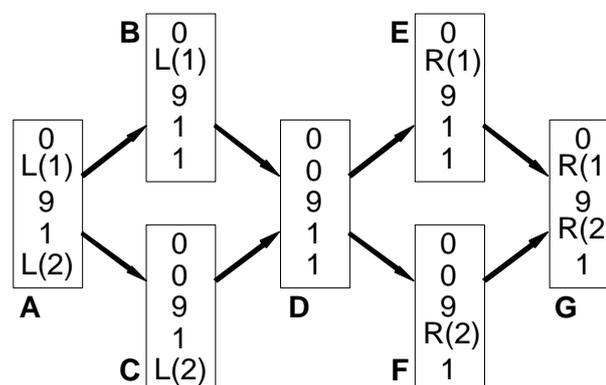

Figure 3. A state-transition network for the phoneme /e/ in the utterance /ten/.

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).

Before I go on to the next section, in which the conventional and new HMMs are compared, I would like to address another important aspect of the new HMMs, namely data sharing. In Figure 4 the networks are shown for the /e/ in /pen/ and /den/, respectively. A comparison of the networks for /ten/, /pen/ and /den/ reveals that many of the states in these networks are similar. Only the states A of /ten/, and A and B of /pen/ occur once in this example, while the states D, E, F and G are the same in the 3 cases. The consequence is that in this way data sharing is possible. The data sharing obtained in this way is based on an articulatory framework, i.e. data are shared across states in which the articulatory gestures are expected to be similar. Clearly, the choice of the feature set is important in this respect.
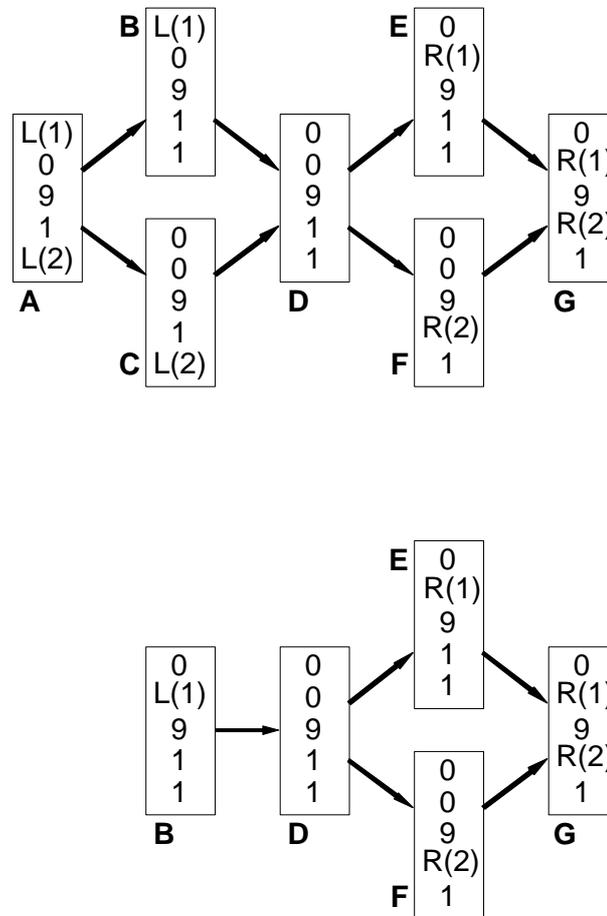

Figure 4. State-transition networks for the phonemes /e/ in the utterances /pen/ and /den/, respectively.

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, articulatory 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 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 about articulation 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. Furthermore, it will also enlarge our understanding of speech production.

Acknowledgements

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 théorie unifiée. In: Bernard Laks & Annie Rialland (eds.) Architecture des Représentations 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.