AN INFORMATION THEORY APPROACH TO SPEAKER ADAPTATION

Gerhard Rigoll
Fraunhofer-Institut (IAO)
Stuttgart, West Germany

ABSTRACT

This paper describes a novel approach to speaker adaptation. The work was carried out by the author while he was a visiting scientist at the IBM Thomas Watson Research Center in Yorktown Heights/USA. The purpose of the research was to train the IBM speech recognition system with only five minutes of speech and to obtain at least a 95% recognition rate after adaptation for a 5000 word vocabulary recognition task. The adaptation algorithm is based on an Information Theory approach used for estimating the label stream of the new speaker by using a stochastic model describing the spectral differences between the new and a reference system. During an evaluation where twelve speakers were tested in ordinary training mode the average recognition rate for a 5000 word vocabulary was 96.4%. When the speakers were tested in 5 minutes adaptation mode the recognition rate dropped to 95.2%. A very important point is that the average decoding time increased by a factor of 1.35 while this factor is often 3-5 if other adaptation algorithms are used.

INTRODUCTION

Most large vocabulary speech recognition systems operate in speaker-dependent mode. Since large vocabulary speaker-independent speech recognition is still a very difficult task, speaker adaptation has been a very interesting topic during recent years and several speaker adaptation algorithms have become very popular (4)-(6). The IBM speech recognition system is a large vocabulary speech recognition system based on Hidden Markov Models and was the first speech recognition system of this kind (3). It operates with a 5000 or 20000 word vocabulary in speaker-dependent mode. The system is trained with 100 sentences uttered by the new user. These 100 sentences represent approximately 20 minutes of speech from that user. The speech data is transformed by the use of signal processing algorithms into a sequence of feature vectors containing ear model parameters. 200 prototypes are derived from the first five minutes of speech (3). The speech data is then further transformed into a label stream by assigning each feature vector to the closest prototype. The label stream representing the speech data of the 100 sentences is the input to the Forward-Backward algorithm that is used for estimation of the Markov Model transition and output probabilities.

It would be a great advantage if it was possible to reduce the amount of training data substantially. Unfortunately, in this case the data is no more sufficient in order to estimate the parameters of the Hidden Markov Models of the speech recognition system. Therefore a speaker adaptation algorithm has to be used that takes the short training script of the new speaker and combines these data with speech data obtained from one or more reference speakers for which the full 100 sentences training script was available. Fig. 1 shows how the standard large vocabulary adaptation algorithms work. The normal training procedure for a Hidden Markov Model based speech recognition system is shown on the left side of Fig. 1. The short training script of the new speaker shown on the right side of Fig. 1 is first processed in the same way as the speech data used for training the system. The Markov Model parameters obtained from normal training with the reference speaker are now modified using the information obtained from the short training script of the new speaker. A probabilistic spectral mapping based on a confusion matrix has proved to be very useful for that purpose (4).

Fig. 1: Basic principle of the standard adaptation algorithms

If such an algorithm is applied to the IBM speech recognition system, this can lead to a satisfactory recognition rate after adaptation but the time necessary for decoding a sentence will be very long, leading to a very slow speech recognition system. This is due to the fact that the output probabilities of the Markov Models obtained from the reference speaker were smoothed out during the adaptation described in Fig. 1. The statistics for the new speaker are not very sharp and the stack decoding algorithm used in the IBM speech recognition system will search many different paths until the path yielding the highest probability is found. That is the reason for the long decoding times. The goal for developing a new adaptation algorithm for the IBM speech recognition system was therefore as follows: reducing considerably the amount of speech data necessary for training while maintaining a recognition rate of at least 95% for a 5000 word vocabulary speech recognition task and taking into account only quite a small increase of the decoding time.
COMPLEXITY OF THE ADAPTATION TASK

The complexity of the adaptation task is mainly determined by three factors: the size of the vocabulary for the recognition task to be performed after adaptation, the required recognition rate after adaptation and the amount of speech that is available for the new speaker of the system. If it is desired to use only a few seconds of speech data from the new speaker the adaptation will be very fast and easy for the new user, but for a large vocabulary, as in the case of the IBM speech recognition system, this may not result in a very high recognition rate. It is expected that the new user would rather like to spend a few minutes of training with the system in order to obtain a recognition rate of at least 95% which would make the speech recognition system still attractive for his purpose to use it as a dictation machine. It was therefore decided to use five minutes of speech from the new speaker and to require at the same time that the error rate of the system should not increase by more than 1-2% after adaptation. Since 5 minutes are usually evaluated by the system in order to derive new prototypes for the speaker, it would be possible in this case to perform adaptation using the personal prototypes of the new speaker. What these conditions mean for the IBM speech recognition system can be seen in Fig. 2 showing the recognition rate of that system if trained with various amount of speech data.

The normal 20 minutes training of the system results in a recognition rate of 96-98%, depending on the investigated speaker. It is therefore amazing to see that a five minutes training of the system usually leads to a recognition rate above 90%. It has also turned out that if no adaptation is used at all and the new speaker simply uses the Markov Model parameters of the old speaker, a recognition rate close to 90% can sometimes be obtained. That shows that it is not very difficult to reach a recognition rate of about 90-92% with any kind of adaptation method for that system. The real difficulty is to obtain recognition rates of 95% or more after adaptation. It can be seen from Fig. 2 that in the speaker-dependent case it takes 15 more minutes of the speech data of the original speaker to increase the recognition rate from 92% to 96%. This makes the complexity of the adaptation task more obvious, because in this case, the 15 original minutes bringing the system from 92% to 96% are replaced by 20 minutes of speech data from a completely different speaker to increase the recognition rate from 92% to 95%.

DESCRIPTION OF THE ADAPTATION ALGORITHM

The application of Information Theory algorithms did result in very powerful speech recognition systems as e. g. described in (2)-(3). Therefore, the attempt was made to apply Information Theory principles also to the problem of speaker adaptation. The basic idea of the Information Theory approach in speech recognition is the assumption that the information of the words of a spoken sentence is coded in the speech signal. If it is possible to build a model describing the coding scheme then one can design a decoder capable of extracting the word sequence back from the speech signal and reconstructing the sentence in that way. Therefore, the principle question here is: how can one use the information provided by the speech signal in order to reconstruct some data (in this case the word sequence) that has generated this signal? If one tries to apply the Information Theory principle to speaker adaptation one could ask the very similar question: how can one use the information provided by the reference speaker (i. e. his speech data) in order to reconstruct some data originating from the new speaker? If that is possible, then one has some data available that characterizes the new speaker and this data could then be used for training the recognition system similar to the way the speech signal of the reference speaker is used for training the system. It is therefore very useful to define the label stream that the new speaker would produce for training as the data that is desired to be reconstructed from the speech signal of the old speaker by applying the Information Theory algorithm. This is shown in Fig. 3 where the Information Theory principle applied to speech recognition and speaker adaptation is compared.

In the first case, it is assumed that the sentence $W$ to be recognized has been produced by a sentence generator. The speech recognition system has no access to the generated sentence and can only process the label stream $Y$ that is imagined to be generated by the acoustic channel. If it is possible to model the acoustic channel then one can design a linguistic decoder that picks the word sequence $W$ that maximizes the probability $P(W|Y)$. In the case of a speech recognition system, a stochastic model is established for the acoustic channel and the decoder is based on the Viterbi algorithm.

For speaker adaptation, one imagines that the new speaker has produced the label stream $Y$ necessary for training the system. Unfortunately, one has no access to that label stream and only the speech data of the reference speaker is available for processing. This data is represented by a sequence of feature vectors denoted as $X$. If it is possible to model the relation between the label stream $Y$ of the new speaker and the speech data $X$ of the old speaker,
interpreted as "channel" between old and new speaker, then one can design a decoder that reconstructs the label stream Y maximizing P(Y/X).

It is then possible to use that reconstructed label stream for training the speech recognition system in order to obtain the parameters of the Hidden Markov Models describing the new speaker. The same approach as in the case of speech recognition was chosen for the realization of the channel model and the decoder. That means that the relation between the old and the new speaker has been modelled with the use of a Markov Model shown in Fig. 4 and the decoder is based on the Viterbi algorithm.

Fig. 4: Markov Model used for describing the spectral differences of two speakers.

The states of that Markov Model correspond directly to the 200 prototypes of the new speaker where transitions between all states are possible. The output of the Markov Model is a Gaussian distribution of the twenty ear model parameters the old speaker would produce if the new speaker produced a certain prototype at the same time. The model can be trained by looking at the label sequence resulting from the short training script of the new speaker and the corresponding acoustic data produced by the reference speaker if he utters the same sentences. For that purpose, the 5 min. of speech data available for the new speaker is signal processed and prototypes are derived that are used for labelling that data. The first five minutes of the 20 min. speech data from the reference speaker are time aligned to the five minutes of the new speaker using dynamic programming. By looking at the label stream of the new speaker it is possible to calculate the transition probabilities for the Markov Model from Fig. 4. By looking at the time aligned acoustic data of the reference speaker it is possible to calculate the corresponding output distributions. This model indeed describes the spectral differences between the two speakers since it predicts which acoustic prototype Y is produced by the new speaker if the old speaker produces a certain parameter vector \( \mathbf{X} \). It is therefore also possible to interpret this approach for speaker adaptation as an approach based on the use of a learning system that is able to learn the spectral differences of the two speakers.

Once the model has been obtained it is possible to use the Viterbi algorithm as decoder in order to reconstruct the label stream of the new speaker. The Viterbi algorithm takes the acoustic data \( \mathbf{X} \) of the reference speaker and calculates the optimal state sequence Y of the underlying Markov Model from Fig. 4 by maximizing P(Y/X). The state sequence Y is approximately the label sequence the new speaker would have produced if he had spoken all the training sentences. Thus, the acoustic data of the old speaker has been transformed into the label stream of the new speaker. That label stream and the label stream resulting from the 5 min. of original speech from the new speaker are then used for training the system with the standard Forward-Backward algorithm in order to estimate new Markov Model parameters for the new speaker. The entire adaptation scheme is outlined in Fig. 5 which can be compared to Fig. 1.

Fig. 5: Outline of the new adaptation algorithm

The approach has the following advantages:

- The output probabilities of the Markov Models for the speech recognition system are not smoothed out as described in the introduction, because these probabilities are not derived from old values in this case. This results in much shorter decoding times.
- It is easy to recombine the reconstructed 20 minutes label stream with the original 5 minutes label stream obtained from the new speaker by simply appending the streams. As mentioned before, the original 5 minutes data of the speaker are very valuable.
- The algorithm leads not only to modified output probabilities of the Markov Model but also to modified transition probabilities.
- Only one single reference speaker is necessary for adaptation and it is indeed more advantageous to concentrate on modeling the spectral differences between that single reference speaker and the new speaker rather than attempting to use multi-speaker information.
- It is possible to model the spectral differences of the two speakers using very personal data for both of the speakers, namely the original prototypes of the new speaker and the original feature vectors of the old speaker. A confusion matrix approach describes which prototype the new speaker generates if the old speaker generates a different prototype, but the prototypes will be only personal for one of the two speakers. A model using data which is personal for both speakers should be more suitable.
- The Information Theory principle can now be used consequently for speech recognition and for speaker adaptation.

The disadvantages are:

- The algorithm requires rather long computing times mainly because the time consuming Forward-Backward algorithm has to be applied
after the label stream has been reconstructed.

There are certain problems in selecting the best reference speaker out of a set of possible candidates, although this is a general problem for every adaptation algorithm that is not based on a multi-speaker approach. The ways for selecting speakers and the possible problems are quite complex and a few more details are mentioned in (1).

EVALUATION

The algorithm was tested in the following way: 12 speakers were trained in ordinary speaker-dependent training mode and were tested on a 5000 word vocabulary recognition task consisting of 50 sentences, which is usually used for testing the performance of the IBM speech recognition system. Then each speaker was tested in adaptation mode by finding for each speaker the best reference speaker out of the set of the 11 remaining speakers and training the system using only 5 minutes of the original speech data and 20 adapted minutes of the selected reference speaker. The results can be seen in Table 1. It can be seen that the average recognition rate was 96.4% for normal speaker-dependent training and that this rate dropped to 95.2% if adaptation was used. The average CPU time increased from 21 to 29 minutes. Some interesting cases are speakers 7 and 10 where change of the recognition rate was only very small. Very interesting is speaker 12 who was the only female speaker and was very well adapted using the data of the male speaker No. 2. Speaker No. 4 had a foreign accent and it can be seen that his recognition rate dropped more than in average but is still acceptable after adaptation. Only five speakers out of the 12 speakers were found to be useful reference speakers.

CONCLUSION

It was possible to show that the application of Information Theory principles to the problem of speaker adaptation seems to be useful and leads to acceptable recognition results. Within the field of Information Theory there are now many refinements of that approach possible that could lead to a more sophisticated use of that principle and therefore may lead to improved adaptation algorithms in the future.

REFERENCES


