A CONFIDENCE MEASURE FOR DETECTING RECOGNITION ERRORS IN ISOLATED WORD RECOGNITION

Olli Viikki, Kari Laurila, Petri Haavisto

Nokia Research Center
Speech and Audio Systems Laboratory, Tampere, Finland

ABSTRACT - Error detection is an important technology needed to improve the usability of practical speech recognition systems. In this paper, we propose a confidence based error detection approach for isolated word recognition using Hidden Markov Models (HMM). An on-line garbage modelling technique is used to obtain the reference score for the recognition result. The confidence is defined as difference of the recognized word model score and the garbage model score between the recognized utterance endpoints. Experiments indicate that a large number of recognition errors can be detected even using a small rejection threshold. In the clean environment, we are able to detect and reject over 80% of incorrect recognitions without rejecting any correct recognitions. If the speech signal is corrupted by background car noise, over 60% of recognition errors can be rejected and, 95% of correctly recognized utterances are still accepted. Experiments also show that the proposed technique is capable of rejecting out-of-vocabulary words.

INTRODUCTION

The implementation of practical speech recognition systems have created the need to develop techniques for detecting recognition errors. Current isolated word recognition systems are capable of achieving a very high recognition accuracy in laboratory conditions. In practice, the limited amount of training data and the mismatch between training and testing environments nevertheless slightly reduce the recognition rate. Although the system generally obtains a high recognition accuracy, there are always occurring some incorrect recognitions that can make it difficult to use the application. If the initial recognition rate is close to 95%, it is very difficult to obtain significant recognition accuracy improvements using the current state-of-the-art HMM based speech recognition technology. Therefore, in this paper, we have selected another approach to further reduce the occurrence of recognition errors and improve the performance of the practical speech recognition system. We do not directly try to improve the recognition accuracy, but we present a technique that effectively detects incorrect recognitions during the recognition phase, when the usability of speech recognition systems increases.

In general, three type of recognition errors can be distinguished: deletion, substitution, and insertion errors. In the case of deletion error, the user says the utterance, but the system does not recognize anything. Substitution error occurs when an incorrect word is recognized instead of the correct one. Insertion error corresponds to the case where an additional word is recognized, even though the user has not said anything. From the user point of view, it is obvious that substitution and insertion errors are the most annoying ones, since they always cause some invalid operations to a system, whereas in the case of deletion error the user must just repeat the previous utterance. The objective of this paper is to detect potential substitution and insertion errors already during recognition and change them to be deletion errors. This allows us to avoid a number of invalid operations of a system that would have caused due to substitution or insertion error.

The error detection scheme described in this paper is based on the confidence measure representing the reliability of the obtained recognition result. For each recognition result, a confidence value describing the goodness of the match between the uttered and recognized word is computed, and depending on the confidence value, the recognition result is either accepted or rejected. If the confidence of the recognized utterance falls below a pre-determined threshold value, the recognition result is rejected and the user is asked to repeat the last speech action. A low confidence value of recognition can result from two different reasons. At first, the uttered word does not belong to active vocabulary (out-of-vocabulary word - insertion error), or secondly, the match between the speech input

and recognized word is poor (substitution error). The selection of an appropriate threshold value is critical for the success of the rejection scheme. If the rejection threshold is set to too low, then a large number of incorrect recognitions are accepted, whereas in the case of too high threshold value, in addition to recognition errors, a number of valid recognitions are also rejected. As the recognition accuracy requirements change depending on the application, the optimal threshold value must be adjusted individually for each system. We show that a large number of incorrect recognitions can be detected and rejected without deleting too many correct recognitions. Hence, the suggested error detection technique helps us to develop user-friendly speech recognition systems. This confidence based rejection can be seen as a post-processing block of a recognition system that checks the validity of the recognition result. In this paper, deletion errors are not regarded as normal recognition errors, but they are considered as the rejection of the utterance associated with the low confidence value. This assumption enables us to calculate the improved recognition accuracy, when all low confidence recognitions are rejected.

CONFIDENCE ESTIMATION

In general, a speech recognition system provides a score to characterize how well an HMM matches to an unknown utterance. This score is nevertheless relative, and it depends very much on the speaker and usage environment. Hence, it cannot be directly used to measure the goodness of the match. In order to detect substitution errors or out-of-vocabulary words we need a reference score s_{ref} to test if a recognized word really exists in the input speech segment. By comparing the score provided by the HMM recognizer and the reference score, we can estimate how good the match between the utterance and HMM is. The fundamental requirement for this reference score is that regardless of input speech, it must be quite close to the score produced by the HMM of the recognized word. Now, the confidence C can be defined as distance between these two scores as

$$C = s_{HMM} - s_{ref} \tag{1}$$

As both scores are determined between the same utterance endpoints, we do not have to perform any time normalization, but the scores can be directly compared to each other. Normally, in correct recognitions there is a significant gap between these two scores, and correspondingly, in the case of misrecognitions, i.e., substitution or insertion error, these scores are very close to each other. By selecting an appropriate threshold value, we are able to separate correct and incorrect recognitions from each other based on their confidence score.

Reference Score Computation

There are a number of different ways to compute the reference score. The issue in the reference score determination is similar to that of keyword spotting, namely, we must define a model whose log-likelihood score is, in the case of valid utterance, less than that of recognized HMM, but in the presence of out-of-vocabulary word or substitution error, the score of this model is greater than the score of the recognized word HMM. In (Rahim *et al.*, 1995) two different approaches to obtain the reference score for utterance verification are presented. Both the methods are based on the use of acoustic filler models that represent out-of-vocabulary words. In the first method, a digit independent general acoustic filler HMM was trained using all non-keyword speech, whereas in the second approach an additional keyword-specific anti-HMM was estimated. The major drawback of these two approaches is the model parameter estimation. It is extremely difficult to adjust the model parameters for the filler HMMs so that an acceptable out-of-vocabulary word rejection rate is achieved in all possible noise conditions with all speakers. In addition, in the latter case the filler models are vocabulary specific, when a new filler HMM must be estimated for each new vocabulary.

An alternative solution to keyword spotting is presented in (Boulard *et al.*, 1994) where one does not attempt to explicitly train any filler HMM using general speech, but so called garbage model is used to characterize out-of-vocabulary words. The basic idea behind garbage modelling is that the score of the individual speech frame is never the best one, but it is always one of the top candidates. This means that the utterance is recognized using garbage model, only in such cases, when the match between

the utterance and keyword HMMs is poor. For each speech frame, the garbage model score is computed as the average of the n best scores produced by keyword HMMs. Therefore, this technique is often referred to as "on-line garbage modelling". Both the on-line garbage modelling technique and acoustic filler HMMs can also be applied to the detection of substitution and insertion errors. In this paper, we show how on-line garbage modelling can be applied to detect recognition errors. The reference score s_{ref} needed in the confidence estimation is computed by means of the on-line garbage model score.

Confidence Score Computation Using On-line Garbage Modelling

As shown in Equation (1), the confidence estimate computation requires two different scores. The score of the recognized model s_{HMM} and the reference garbage model score s_{ref} are determined between the recognized utterance endpoints as shown in Figure 1.

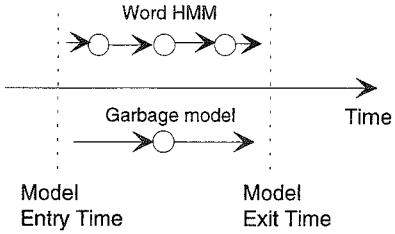

Figure 1. Scores required for confidence computation

Compared to the garbage model score computation method presented by Boulard *et al.*, we use a slightly modified approach. For each speech frame, the maximum log-likelihood s_{max} and the average log-likelihood s_{ave} are searched. Now, the local garbage model score at each time instant i with the desired rank value r is calculated using a linear interpolation between the maximum and average scores. The rank value r is restricted so that the average score corresponds to the rank value 0.50 and the maximum score the rank value of 1.0, respectively. By changing the rank value, we can adjust the garbage model output score. From the implementation point of view, this type of garbage model score computation method is more attractive than the method presented by Boulard *et al.* The HMM based recognizer defines the endpoints of the utterance and the local on-line garbage model scores are then accumulated between the start and end points of the utterance. Thus, the reference score needed for the confidence calculation can be expressed as

$$s_{ref} = \sum_{i=EntryTime}^{ExitTime} s_{garbage}(i) \qquad (2)$$

The basic assumption in the confidence scheme is that in correct recognitions, the states of the recognized HMM produce greater scores for almost all speech frames than other HMMs. Then, the accumulated score of the recognized HMM is greater than that of the garbage model. In the case of incorrect recognition, the states of the recognized HMM do not continuously produce the highest likelihood scores for all speech frames, but the maximum scores for each speech frame are distributed over several HMMs, when the accumulated score of the garbage model is close (or over) to the corresponding score of recognized HMM.

BACKGROUND MODELLING USING GARBAGE MODEL

It is also possible to use the garbage model to recognize background silence/noise instead of using conventional HMM whose parameters are estimated from the noise regions of training utterances. For background modelling, it is useful that one can adjust the garbage model output score by changing the rank value. In (Iso-Sipilä *et al.*, 1996) it was shown that the optimal rank value in terms of recognition

accuracy depends on noise conditions. In our case, however, the constant rank value was used, since in isolated word recognition, one cannot modify the distributions of deletion and insertion errors by means of background models, as in connected or continuous speech recognition. In our system, no conventional HMM based background modelling was used, but the same on-line garbage modelling technique, which was needed in confidence determination, was also used for background noise recognition. For background modelling, a slightly smaller rank value was used than for confidence determination.

TEST DATABASE AND EXPERIMENTS

The test database used in all experiments consisted of isolated first names spoken by five different Finnish male speakers in a noise-free environment. In total, there were 30 names active in the vocabulary to be recognized. Each speaker uttered all names twelve times. One repetition of each word was used for the HMM estimation. The FFT-based mel-weighted cepstral coefficients, their first- and second-order time derivatives, and energy terms were extracted from the continuous speech signal. Left-to-right, state duration constraint (Laurila *et al.*, 1996), speaker-dependent HMMs were estimated using only a *single* training utterance. In all experiments, the garbage model rank for background modelling was set to 0.80 and for confidence calculation the rank was set to 0.90, respectively.

Detection of Substitution Errors

The purpose of these experiments was to detect substitution errors during recognition. Results presented in this context were derived averaging the performance obtained on each of five speakers. In the baseline test, where all recognitions were accepted regardless of their confidence value, the recognition accuracy was 94.4%. Figure 2 shows how the recognition rate increases as the function of the threshold value when the confidence based rejection scheme was applied to isolated name recognition. It should be noted that even using a negative threshold, i.e., the case where the garbage model score is greater than the actual score produced by the recognized HMM, a large number of incorrect recognitions can be detected.

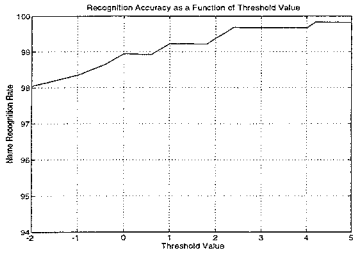

Figure 2. Recognition rate in clean environment

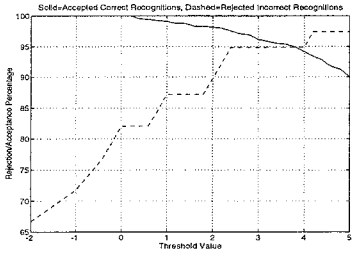

Figure 3. Rejection of correct and incorrect names

In Figure 3, it is shown how the percentage of correctly (recognition error) and incorrectly rejected (correct recognition) utterances change depending on the threshold value. One can see that using a small positive threshold value, we are able to detect over 80% of recognition errors without rejecting any correct recognition. By selecting a higher threshold value, the number of recognition errors can be further decreased. However, it must be noted that the use of high threshold value tends to reject some correct recognitions as well.

As practical speech recognition systems are often used in the presence of noise, it is important that the proposed confidence scheme provides a satisfactory performance, when speech is corrupted by background noise. Noisy utterances were obtained by adding artificially car noise to clean utterances so that the Signal-to-Noise Ratio (SNR) was around 2 dB. However, exactly the same speaker-dependent HMMs were used as in the previous experiments with clean input speech. Figure 4 shows

the recognition rate improvement due to the confidence scheme, and in Figure 5, it is presented how the shares of correct and incorrect rejections change as a function of the threshold value. In the baseline test, the name recognition rate was 90.2%. By introducing a small rejection threshold value, a number of substitution errors can be removed. It should be noted that in the presence of noise, the rejection scheme does not work as efficiently as in the case of clean input speech. By studying Figure 4, one can see that the recognition accuracy increase is not so rapid as in the clean environment. Furthermore, the accepted correct recognition curve falls steeper than in the noise-free environment. The reason for this phenomenon is that noisy utterances are more or less distorted, hence, their general confidence level is lower than that of clean utterances. However, even in the presence of noise, the majority of recognition errors can be detected using a small threshold value in such a way that almost all correct recognitions are accepted.

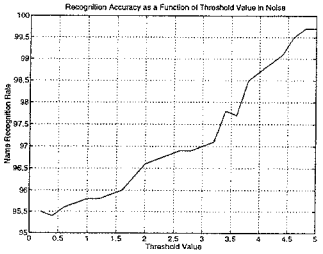

Figure 4. Name recognition rate in noise

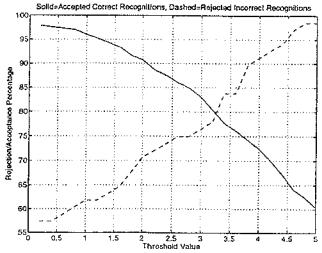

Figure 5. Rejection rates in noise

Detection of Insertion Errors

A speech recognition system should also be capable of detecting words that do not belong to the active vocabulary. The proposed confidence scheme can also be applied to out-of-vocabulary word detection. In the case of out-of-vocabulary word, there is no HMM that would model the input utterance. Therefore, it is very common that the local score peaks for each speech frame are produced by different HMMs, and the measured confidence is reasonably low. It must be noted that the rejection performance of out-of-vocabulary words cannot be determined explicitly. By choosing out-of-vocabulary words to be tested so that they are close to valid vocabulary words, a large number of insertion errors is occurred. Correspondingly, a very high out-of-vocabulary word rejection rate can be achieved if the word to be tested differ much from valid vocabulary words.

To test the insertion error detection capability, we divided the vocabulary used in the earlier tests into two sets, each containing 15 words. The first set formed the valid vocabulary and the second set was used to test how often out-of-vocabulary words cause insertion errors (false alarm). Figure 6 presents the dependency between the out-of-vocabulary word rejection rate and the threshold value. If no confidence based rejection was used, all out-of-vocabulary words were recognized as a valid vocabulary word.

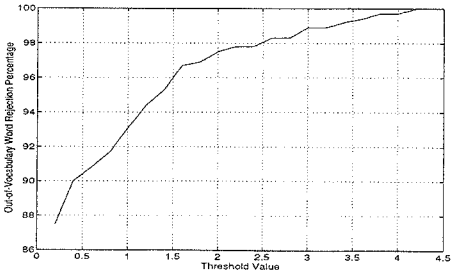

Figure 6. Out-of-vocabulary word rejection

CONCLUSION

In this paper, a method for detecting recognition errors and out-of-vocabulary words was introduced. For each recognition result, a confidence value is determined as distance between the recognized HMM score and garbage model score. Such recognitions whose confidence level is below an empirically chosen threshold value are rejected. From the implementation point of view, the proposed technique is attractive, since any additional HMMs need not be stored, and the computational overheads associated with the on-line garbage modelling are minor. The performed experiments show that the majority of recognition errors can be deleted without rejecting correct recognitions too frequently. Thus, a number of invalid operations of a speech recognition system can be avoided if the validity of recognition result is checked using the proposed error detection technique. In the presence of noise, the recognition result reliability is not as high as in the clean environment, and therefore, the recognition accuracy increase is not so dramatic as in the noise-free environment.

REFERENCES

Boulard, H., D'hoore, B., Boite, J.-M. (1994) *Optimizing recognition and rejection performance in wordspotting systems,* Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I-373 - I-376, Adelaide, Australia.

Iso-Sipilä, J., Laurila, K., Haavisto, P. (1996) *Optimal adaptive garbage modelling in speech recognition*, Proceedings of the IEEE Nordic Signal Processing Symposium, Helsinki, Finland.

Laurila, K., Majaniemi, M., Yang, R., Haavisto, P. (1996) *Speech recognition with state duration constrained Maximum Likelihood HMMs*, Proceedings of the IEEE Nordic Signal Processing Symposium, Helsinki, Finland.

Rahim, M.G., Lee, C.-H., Juang, B.-H. (1995) *Robust utterance verification for connected digits recognition*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 285- 288, Detroit, USA.