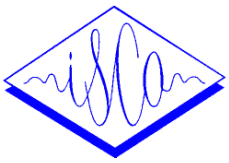

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

WORD REJECTION USING MULTIPLE SINK MODELS

Carlos J. Teixeira and Isabel M. Trancoso[1]

INESC / IST
INESC, R. Alves Redol 9, 1000 Lisbon, Portugal

ABSTRACT

The purpose of this work is two-folded: to improve the robustness of a baseline isolated-word recogniser with a medium-sized vocabulary in terms of word rejection capabilities and to verify these improvements on a multi-application database including different native and non-native English accents.

Although the results obtained so far could be biased by the unavailability of realistic corpora, they seem to indicate the usefulness of multiple sink models in this context, relative to the use of a single one. The improvements are particularly evident in multi-accent environments, where our results show that merging two different accents in the training material yields scores which are similar to the ones obtained in a single accent environment.

1. INTRODUCTION

One of the main purposes for having automatic speech recognition (ASR) is to provide a more human-like man-machine interface communication. For most applications of ASR, it is generally assumed that users speak only the predefined vocabulary words in isolation. However, this is not a realistic assumption, as many users will speak the vocabulary items along with extraneous speech input. In an application including, for instance, *main menu* and *wake up* as keywords, an untrained user could say *back to the main menu please* or *give me the wake up option*. Although those extraneous words usually increase the robustness of the communication between humans, the present generation of ASR's is not generally able to take advantage from that additional information. The extraneous words produce recognition errors that generate wrong commands, thereby decreasing the recognition accuracy and the robustness of the system.

The purpose of the present work is two-folded: to improve the robustness of a baseline isolated-word recogniser with a medium-sized vocabulary in terms of word rejection capabilities and to verify these improvements on a multi-application database including 5 different non-native English accents. Although the experiments performed so far cover only one specific application and two different accents, the goal is to derive general rules for guiding the training and testing of a set of applications, including different types of accent.

Word rejection is equivalent to distinguishing between patterns belonging to two different classes: the keywords for a specific application, and the remaining words of that particular language (or a set including the most probable extraneous words in a normal testing session). One of the most common solutions for integrating word rejection capabilities in a HMM (Hidden Markov Model) based recogniser consists simply of adding new models/patterns to the existing HMM's. These models, which are trained on the basis of words outside the vocabulary (non-keywords) are supposed to be general enough to produce a higher likelihood for non-keywords in comparison with keywords. While a keyword model (KM) is trained with utterances from a single word, a non-keyword model (commonly designated as Garbage or Sink Model - SM) is trained with utterances from several words in order to be representative not only of those words but of any non-keyword.The successful use of SM's for word rejection in small vocabulary applications (5-25 keywords) has been reported by several authors [1,2]. For larger vocabularies, however, the rejection task is more difficult, as the probability that a non-keyword can reach a higher likelihood value with a KM than with one of the fewer SM's increases.

The eventual use of a recognition application in the pan-European telephone network using, for instance, the English language brings up new questions about how different native and non-native accents can affect the performance of the recogniser. One general procedure to deal with this problem consists of viewing it as a speaker independence problem and building a training set which covers every type of speaker variability. The collection of such a training database, however, is a very consuming task. In order to have a preliminary measure of this specific problem, we have used two databases with the same vocabulary but using different English accents from two European countries.

A HMM-based recogniser, developed in the scope of ESPRIT Project SUNSTAR [3], was used for the experiments described in this paper. The built-in endpoint detector used for extracting isolated words closely follows the energy-based algorithm described in [4]. A linear ten-state topology was used for each model with no skips over the states.The output probabilities are described by a single Gaussian mixture with a diagonal covariance matrix. The speech signal is sampled at 8 kHz. Each observation is computed over a 20 ms Hamming window with 50% of overlapping. The observation vectors include the cepstrum and delta-cepstrum coefficients. The cepstrum is based on an 8th order LPC analysis.The delta-cepstrum uses a delay of four frames.The training phase is initialised with a Viterbi forced alignment. The Baum-Welsh algorithm is then iteratively computed until the overall likelihood falls bellow a threshold or a maximum number of iterations is exceeded. In the present work, no grammar was used in order to keep the results as general as possible.

Section 2 of this paper describes the speech corpora used on the following experiments and its selection from the SUNSTAR multi-national speech databases. The base-line recognition and rejection scores obtained within each speech corpora accent and the effects caused by using a different training and testing accent are described in Section 3, together with the use of multiple SM's. In Section 4, different models set-ups, including single and multiple SM's, are used in an attempt to achieve comparable base-line performance with the double accent testing material. Finally, Section 5 presents our conclusions and guidelines for future research in this promising area.

2. SPEECH CORPORA

A total of five European languages are represented in the SUNSTAR multi-language speech database. Four industrial partners from Denmark, Italy, Germany and Spain each had specific applications to implement and demonstrate. It was decided that all applications should be demonstrated in both their native language and in English, the latter corresponding to translations of the vocabularies used in the respective applications.

Since the demonstrations were planned to be carried out multi-nationally, it was decided to use both native and non-native English speakers for the recording of the corresponding speech corpora, in order to obtain a representative multi-accent English database. 20 speakers from each country and 20 English speakers were selected

[1]representing the SUNSTAR consortium.The work in the SUNSTAR project is done within the framework of the ESPRIT programme and partly funded by the Commission of the European Communities. The following companies form the SUNSTAR consortium: Jydsk Telefon (DK), to which INESC (P) and Speech Technology Centre (DK) are associated, Alcatel FACE (I), Fraunhofer Gesellschaft/IA (D), and Telefonica I+D (E).

10.21437/ICSLP.1992-385

for recording each application. Thus, for instance, a translated German application was recorded using 20 German, 20 Danish, 20 Spanish, 20 Italian and 20 native English speakers. The spoken English corpora is therefore the largest of the five language corpora available. The experiments described in this paper were restricted to the use of the English corpora recorded by native and Danish English speakers. This speech material was maintained in separated training and testing sets in order to evaluate the effects of different accents in the recognition scores but mainly in the rejection scores. Some experiments were also performed joining the two accents in the same set, in order to evaluate an eventual decrease in performance. For these last experiments, the number of speakers and consequently the number of utterances will be double.

The vocabulary of the Danish application (the English translation from the Danish application) was selected as the keyword vocabulary for the baseline recogniser. The size (40 words) and the vocabulary of this set were considered representative, allowing extrapolation to the other SUNSTAR applications. The speech material used for training the keyword models was uttered by 16 speakers (2 repetitions, total of 1280 utterances). The keyword test set was uttered by 2 male and 2 female speakers (2 repetitions, total of 320 utterances). Word rejection experiments were then performed by investigating the ability to reject the utterances from a complementary part of the database. Only a subset of this part has been used so far for sink model training and testing (10 speakers of the two German applications). The translated Danish and German applications included common single-words (e.g. digits) and composite-words (e.g. *alarm call*) which were therefore excluded from the sink models training set. This set was also divided into training and testing subsets: The first of the two repetitions of 70 words recorded by 6 speakers (420 utterances) was used for training the sink models. The two repetitions recorded by 4 other speakers (2 male + 2 female) of 53 different words were used for testing the rejection capabilities (424 utterances).

3. MULTIPLE VS. SINGLE SINK MODELS

In this section, the performance obtained with multiple SM's is evaluated against the one obtained with a single SM, and a reference number for SM's is selected for the following experiments. Training multiple sink models implies splitting the SM training material into several subsets. Our first experiments involved iterative data-driven splitting procedures [5], k-means and graph-based clustering techniques, using the likelihood values presented by the last reestimation step of the Baum-Welsh algorithm as a similarity measure [6]. The improvements in rejection scores obtained with either these procedures or different HMM clustering techniques, however, were not significant. Our next splitting experiments did not include any information on the training data, other than keeping every utterance of the same word in the same training subset. The adopted procedure used an alphabetical ordered list of the 70 different words used for training sink models. Each of the n subsets sequentially took one word from this list, until the last subset removed its corresponding word. The process started again with the first subset until removing every word from the list.

The first set of experiments described in this paper involved only native English speakers. The number of sink models (n) was varied from 0-7, 10, 14 and 20. The recognition rate was computed as the number of correctly recognised keywords, divided by the total number of uttered keywords, and the rejection rate was computed as the number of correctly rejected out-of-vocabulary utterances, divided by the total number of out-of-vocabulary utterances. The results are illustrated in Figure 1. A slight decrease in the recognition score can be noticed with the increasing number of SM's, but it has no statistical significance. In terms of rejection score, a quasi-linear improvement can be observed from 1 to 5 SM's and a saturation effect after this number. The selected reference number of sink models for the next experiments was therefore five. Except when otherwise stated, all the following tests were made with only 4 speakers of the previous testing set.

Table 1 presents recognition (first 5 columns) and rejection (last 5 columns) scores for 0, 1 and 5 models only. The second and third column for each score represent the lower and upper limits of the 90% confidence interval. The fourth and fifth columns represent the limits for the 95% interval. The 3 characters in the column designated as models indicate the accent of the material used for keyword model training, the number of sink models and the accent of the material used for sink model training, respectively.

Table 1 - Test with native English speakers

Models	Recog.	90%		95%		Rejection	90%		95%	
e0	96.9	94.8	98.1	94.3	98.3	0.0	0.0	0.9	0.0	0.6
e1e	96.6	94.5	97.9	94.0	98.1	62.3	58.3	66.1	57.6	66.8
e5e	96.3	94.1	97.6	93.6	97.8	68.4	64.6	72.0	63.8	72.6

These experiments were repeated for 0,1 and 5 SM's using the set of speech material collected from the Danish English speakers. The results are presented in Table 2. The increase in the rejection rate using 5 SM's instead of a single one is also present here.

Table 2 - Test with Danish speakers

Models	Recog.	90%		95%		Rejection	90%		95%	
d0	96.9	94.8	98.1	94.3	98.3	0.0	0.0	0.6	0.0	0.9
d1d	95.9	93.7	97.4	93.2	97.6	64.9	61.0	68.6	60.2	69.3
d5d	95.3	93.0	96.9	92.4	97.1	71.0	67.2	74.5	66.5	75.1

4. EXPERIMENTAL RESULTS WITH DIFFERENT ACCENTS

4.1. Tests with models trained with a different accent

In order to evaluate how different pronunciations can affect recognition and rejection, the previous trained models were tested with the corresponding set of speech collected from the other speakers nationality. Table 3 show the results obtained with the Danish speakers when using the models trained with the native English speakers.

Table 3 - Test with Danish speakers

Models	Recog.	90%		95%		Rejection	90%		95%	
e0	79.7	75.8	83.1	74.9	83.7	0.0	0.0	0.0	0.0	0.0
e1e	77.5	73.4	81.1	72.6	81.7	59.0	55.0	62.8	54.2	63.5
e5e	76.6	72.5	80.2	71.6	80.9	69.8	66.0	73.3	65.3	74.0

The recognition scores significantly decrease when compared with the results from Tables 1 and 2. This is probably the reason why the improvement in rejection rate from 1 to 5 SM's became statistically relevant. The results obtained with native English speakers when using the models trained with Danish speakers are presented in Table 4.

Table 4 - Test with native English speakers

Models	Recog.	90%		95%		Rejection	90%		95%	
d0	86.9	83.5	89.7	82.7	90.1	0.0	0.0	0.6	0.0	0.9
d1d	79.1	75.1	82.6	74.3	83.2	71.5	67.7	74.9	67.0	75.6
d5d	79.1	75.1	82.6	74.3	83.2	76.4	72.9	79.6	72.2	80.2

Without sink models, the decrease in recognition scores relative to Tables 1 and 2 is not so marked as in Table 3. The inclusion of sink models, however, yields recognition scores which are closer to the previous table. The rejection rate shows an improvement relative with Table 2, which corresponds to the same training sets. The lack of symmetry between Tables 3 and 4, that is, the fact that the keyword models trained by Danish speakers performed better than the ones trained by native English speakers in tests with a different accent, may perhaps be attributed to the greater dispersion of the training material spoken by foreigners.

4.2. Tests with keyword models trained with one accent and sink models trained with a different one

The experiments described in this section use the same keyword and single sink models referred in the previous sections, now combined in a different way. The set of keyword models trained with one of the accents was associated with the single sink model trained with the other accent. The results are included in Tables 5 and 6.

Table 5 - Test with native English speakers

Models	Recog.	90%	95%	Rejection	90%	95%
e1d	96.6	94.5 97.9	94.0 98.1	37.0	33.3 41.0	32.6 41.7
d1e	74.7	70.5 78.5	69.7 79.1	84.7	81.6 87.3	80.9 87.8

Table 6 - Test with Danish speakers

Models	Recog.	90%	95%	Rejection	90%	95%
e1d	66.3	61.8 70.5	60.9 71.2	83.0	79.8 85.8	79.2 86.0
d1e	95.0	92.6 96.7	92.0 96.9	32.8	29.2 36.6	28.5 37.4

The drop in recognition performance across different accents confirms the results from the previous section. However, in the worst previous case when native English trained keyword models were tested with Danish speakers, an additional sharp decrease is now found. In fact, 21.3% of the uttered keywords in this case were now rejected by the SM's, when previously only 5.9% were rejected. This strong effect produced by the SM trained by Danish speakers, was already pointed in section 4.1 and is also stressed by a higher rejection score than the one achieved when the corresponding keyword models were present (Table 3). This last aspect was also found with the native English SM tested with the same accent.

4.3. Tests using a single accent for training sink models and both accents for training keyword models

The mismatch found by using different accents in the keyword domain as well as in the non-keyword domain, was evident in the results presented in the previous section. In order to address the recognition problem prior to the rejection one, we started by training the keyword models with all the keyword material from both accents (64 utterances for each). Another possible attempt would be simply to join the previous sets of keywords models. This approach was eliminated at this stage because, with an average sized vocabulary, doubling the recognition time (if the delay introduced by a smaller number of SM's is not considered) could become critical for real-time implementation. Results using with one of the previous referred single SM are presented in Tables 7 and 8, where the first character in the model designation now indicates a mixture of two accents (m). The recognition decay is no longer evident for any of the accents. However, the rejection rate is rather low if a different accent is used for testing the SM.

Table 7 - Test with native English speakers

Models	Recog.	90%	95%	Rejection	90%	95%
m0	96.9	94.8 98.1	94.3 98.3	0.0	0.0 0.0	0.0 0.0
m1e	95.9	93.7 97.4	93.2 97.6	52.8	48.8 56.8	48.1 57.5
m1d	96.6	94.5 97.9	94.0 98.1	29.3	25.8 33.0	25.1 33.8

Table 8 - Test with Danish speakers

Models	Recog.	90%	95%	Rejection	90%	95%
m0	95.9	93.7 97.4	93.2 97.6	0.0	0.0 0.0	0.0 0.0
m1e	95.0	92.6 96.7	92.0 96.9	23.4	20.1 26.9	19.6 27.6
m1d	95.9	93.7 97.4	93.2 97.6	57.1	53.1 61.0	52.3 61.7

4.4 Tests joining previously trained sink models from each of the two accents

A first attempt to solve the remaining problem from the last section is to join the corresponding SM's trained with each accent. This has not the drawback referred to about the keyword models because of the smaller number of SM's to be joined. Moreover, there are strong evidences that doubling the number of SM's improves the rejection scores. The results are reported in Tables 9 and 10 with 1 and 5 SM's from each accent set.

Table 9 - Test with native English speakers

Models	Recog.	90%	95%	Rejection	90%	95%
m1e1d	95.9	93.7 97.4	93.2 97.6	54.5	50.5 58.4	49.7 59.2
m5e5d	95.6	93.3 97.2	92.8 97.4	62.7	58.8 66.5	58.0 67.2

Table 10 - Test with Danish speakers

Models	Recog.	90%	95%	Rejection	90%	95%
m1e1d	95.0	92.6 96.7	92.0 96.9	60.6	56.7 64.4	55.9 65.2
m5e5d	94.4	91.9 96.2	91.3 96.4	70.6	66.8 74.0	66.0 74.7

The recognitions scores were not significantly affected by the the new SM's. The rejection scores achieved for both accents are similar to those presented in Tables 1 and 2, when the same accent was used for training and testing. It is also important to stress that the improvement found by using 2 X 5 SM's instead of two (2 X 1 SM) becomes now evident with intervals of confidence of 90% for native English speakers and 95% for Danish speakers. This indicates that multiple SM's are particularly useful when different accents are present.

4.5 Tests using both accents for training keyword and sink models

In the following experiments the above referred training speech material collected in England and Denmark were both used for training the same set-up model. A final set of experiments was performed with 1,2,5 and 10 SM's trained by selecting the utterances from both accents (12 repetitions) of the same word, according to the procedure described in section 3. The results are reported in Tables 11 and 12. By comparing with the equivalent results from Tables 9 and 10, namely the experiments including the same number of SM's (models m2m with m1e1d and m10m with m5e5d), no significant changes were found in the rejection rates by mixing the two accents in the same SM's. Here, an improvement in the rejection rate can be also found by using a high number of SM's. The effect is detected with a 95% interval of confidence between using 5 or 10 and using a single SM. The same conclusion is supported between using 2 and 10 SM's. These results also indicate that a higher number of SM's than 5 can provide even better rejection rates when two different accents are present in the training and testing sets.

Table 11 - Test with native English speakers

Models	Recog.	90%	95%	Rejection	90%	95%
m1m	96.3	94.1 97.6	93.6 97.8	48.6	44.6 52.6	43.9 53.3
m2m	95.9	93.7 97.4	93.2 97.6	54.3	50.3 58.2	49.5 58.9
m5m	95.6	93.3 97.2	92.8 97.4	59.7	55.7 63.5	54.9 64.2
m10m	95.6	93.3 97.2	92.8 97.4	64.6	60.7 68.3	60.0 69.0

Table 12 - Test with Danish speakers

Models	Recog.	90%	95%	Rejection	90%	95%
m1m	95.3	93.0 96.9	92.4 97.1	47.9	43.9 51.9	43.2 52.6
m2m	95.6	93.3 97.2	92.8 97.4	55.2	51.2 59.1	50.4 59.9
m5m	95.0	92.6 96.7	92.0 96.9	62.5	58.6 66.3	57.8 67.0
m10m	94.4	91.9 96.2	91.3 96.4	67.9	64.1 71.5	63.3 72.2

5. CONCLUSIONS AND FUTURE DEVELOPMENTS

The two main features of this work which increased the difficulty of the rejection task were the size of the vocabulary and the different accents of the speech corpus. The results presented in this paper were biased by the unavailability of realistic corpora for modeling and testing out-of-vocabulary words. By using application oriented training material, higher rejection rates can be expected. Another limitation concerns the use of a single continuous mixture for the observation vector, which has some influence on the number of SM's which must be adopted. Under the described conditions, the main conclusion from this work is that the use of more than one sink model is useful for increasing the rejection rate. These improvements are particularly evident when testing with different accents, as it was reported with native English and English spoken by Danish speakers. In these multi-accent environments, our results show that merging both accents in the training material yields scores which are similar to the ones obtained in a single accent environment. This can be regarded as an extension of the speaker independent training approach.

Further work is now being done in the connected speech domain including word spotting. The following steps will include experiments using a semi-continuous HMM recogniser and application adaptive SM's. The integration of noise immunity techniques will also be considered covering another important problem of the robust speech recognition area.

ACKNOWLEDGEMENTS

The authors wish to thank Prof. Paul Dalsgaard, Børge Lindberg and Bjarne Andersen from the Speech Technology Centre at the University of Aalborg (Denmark) and Prof. António Serralheiro (INESC-IST). Their help and advice made this work possible.

REFERENCES

[1] J. G. Wilpon, L. R. Rabiner, C.-H. Lee, E. R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", *IEEE Trans. on Acoust., Speech and Signal Proc.*, Vol.ASSP-38, No. 11, pp. 1870-1878, November 1990.

[2] B. Chigier, "Rejection and Keyword Spotting Algorithm for a Directory Assistance City Name Recognition Application", *Proc. ICASSP*, pp. 93-96, March 1992.

[3] SIRtrain Training Software - User Guide, Vers.2.1, *SUNSTAR Esprit Project 2094 Report*, November 1991.

[4] L. Lammel et. al., "An Improved Endpoint Detector for Isolated Word Recognition", *IEEE Trans. on Acoust., Speech and Signal Proc.*, Vol.ASSP-29, No. 4, pp. 777-785, August 1981.

[5] L. R. Rabiner, C.-H. Lee, B. H. Juang, and J. G. Wilpon, "HMM Clustering for Connected Word Recognition", *Proc. ICASSP*, pp. 405-408, May 1989.

[6] C. J. Teixeira and I. M. Trancoso, "Word Rejection Experiments Using the SIRTrain Software on the SUNSTAR Speech Database", *SUNSTAR Esprit Project 2094 Report*, January 1992.

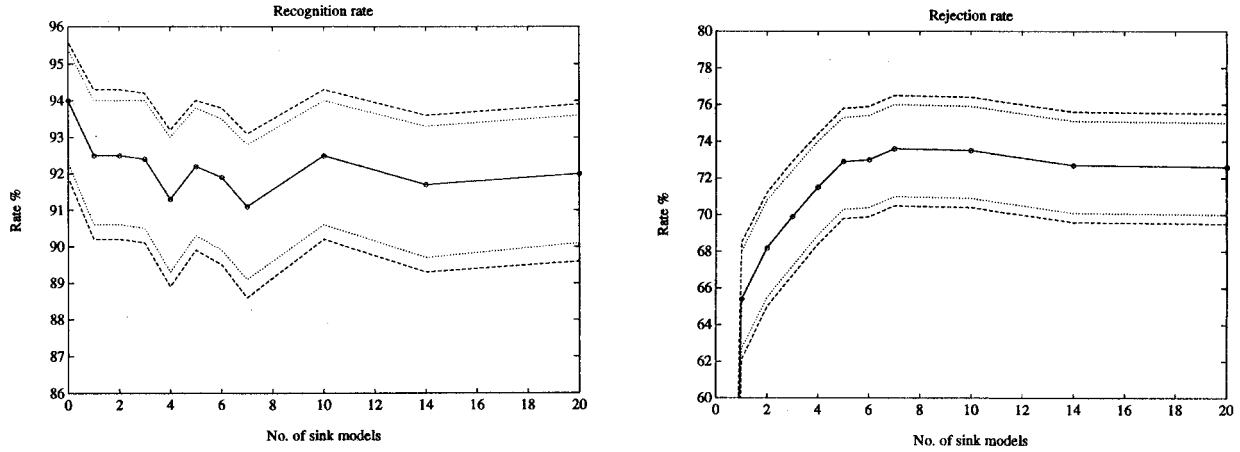

Figure 1. Recognition and rejection rates versus number of SM's for native English speakers. Experimental results are marked with small circles, dashed and dotted lines represent the 95% and 90% confidence intervals, respectively.