INNOVATIVE APPROACHES FOR LARGE VOCABULARY NAME RECOGNITION

Yuqing Gao, Bhuvana Ramabhadran, Julian Chen, Hakan Erdoğan and Michael Picheny

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
email: yuqing, bhuvana, juchen, hakan, picheny@us.ibm.com

ABSTRACT

Automatic name dialing is a practical and interesting application of speech recognition on telephony systems. The IBM name recognition system is a large vocabulary, speaker independent system currently in use for reaching IBM employees in the United States. In this paper, we present some innovative algorithms that improve name recognition accuracy. Unlike transcription tasks, such as the Switchboard task, recognition of names poses a variety of different problems. Several of these problems arise from the fact that foreign names are hard to pronounce for speakers who are not familiar with the names and that there are no standardized methods for pronouncing proper names. Noise robustness is another very important factor as these calls are typically made in noisy environments, such as from a car, cafeteria, airport, etc. and over different kinds of cellular and land-line telephone channels. We have performed a systematic analysis of the speech recognition errors and tackled the issues separately with techniques ranging from weighted speaker clustering, massive adaptation, rapid and unsupervised adaptation methods to pronunciation modeling methods. We find that the decoding accuracy can be improved significantly (28% relative) in this manner.

1. INTRODUCTION

Automatic name dialing is a practical and interesting application of speech recognition on telephony systems. Other researchers [1, 2, 3] in the name recognition area have typically focussed on small vocabulary, speaker dependent and limited application environments. The IBM name recognition system described in this paper is significantly different from those published in the literature in the following ways. It is a large vocabulary (280K names) speaker independent system, and it handles calls from various channels and environments, such as calls from land-line, speaker phone, or cellular, also calls made from office, home, public phone or car, etc. Approximately 10,000 calls are processed each day by this system.

When the list of names to be recognized is as large as 200,000, and when the calls are received from all sorts of different channels and environments, the name recognition task becomes a very complex speech recognition problem. From a practical point of view, the overall accuracy of a name dialing system depends on a number of factors, including the speech recognition accuracy for multiple channels and noise environments, the problem of homonyms and out-of-vocabulary entries in the name list, the search speed when no language model can be used. The overall success rate of the calls received depends not only on the speech recognition accuracy, but also on the call volume handling capability, the response speed and, finally, the user-friendliness of the user interface.

Having described the large vocabulary name recognition problem, it is clear that this cannot be handled in a manner similar to other small or medium vocabulary tasks such as, digit or spelling recognition, or even small vocabulary (less than 100) name recognition applications [1, 2, 3]. Hence, we have approached this problem from LVCSR perspective. New approaches and algorithms are needed in order to achieve high recognition accuracy.

In this paper, we present a systematic analysis of the speech recognition errors and subsequently developed new algorithms which are different from those used in transcription tasks to fit this problem. Specifically, we optimized the context-dependent HMM states and the Gaussians used to model them. We focussed on speaker clustering, rapid adaptation and massive adaptation algorithms and pronunciation modeling to handle the large vocabulary. We also present suggestions for future improvements.

2. SYSTEM DESCRIPTION

The name recognition system is a derivative of the IBM speech recognition system described in [4]. The acoustic models are built from an in-house telephony database of over 600 hours. This data includes spontaneous speech, read speech, personal names, zip codes, business names as seen in the yellow pages, street addresses, credit card and telephone numbers, etc. This data was collected from over 16,000 speakers.

The test data was collected from a live name-dialing system that is currently operational within IBM. Two data sets collected over different periods of time were used. The first dataset, Test.1, contains 5.7K utterances and the second dataset, Test.2, consists of 10.9K utterances. Each utterance includes a first name and a last name, or simply the last name only, or includes the location along with the name, etc. A detailed analysis is presented in Section 3.

The speech recognition system uses an alphabet of 52 phones. Each phone is modeled with a 3-state left-to-right HMM. The acoustic front-end uses a 13-dimensional cepstral feature vector extracted every 10 ms, sentence based cepstra mean normalization and 9-frame spliced LDA. Two different methods, the traditional LDA (lda) and weighted-LDA (wlda) [5], were used to perform linear discriminant analysis.

During recognition, a finite-state grammar is used as the language model. Two different grammars were compiled, one comprising of 8000 unique names (G1) and the other 280K unique names (G2). The lexicon was built with pronunciations derived from linguists and the Random House Unabridged Dictionary [6]. Lexicon L1 consists of 8K pronunciations of 6K words and lexicon L2 consists of 97K pronunciations of 76K words. It should be mentioned here that there was a significant number of foreign names from a wide set of nationalities. Therefore, deriving pronunciations for the varied accents was a monumental task and needed several iterations before use.

The baseline system was decided in the following manner. In an effort to test the variation in the number of states and Gaussians on the recognition error rate, several systems were built using two different decision trees, T1 and T2, both having approximately 3000 context-dependent HMM states. Tree T2 is built off clean training speech data, while Tree T1 is trained from all of the training data, which includes both clean and noisy data. Table 1 presents the error rates obtained from these various systems. From the experimental results, it showed that systems using Tree T2 perform much better than those using T1. Therefore it is clear that for the name recognition task, the decision tree, which models the phonetic context of phones, should be trained using clean data. The best performing system is a lda, 40K Gaussian system built off the decision tree T2, with error rates of 11.60% and 5.45% for Test.1 and Test.2, respectively. Therefore we decided to use this as our baseline system to evaluate the new algorithms proposed.

Decision tree T1			
type of lda	testset	# of Gaussians	
		70K	40K
lda	Test.1	11.77%	13.02%
	Test.2	5.75%	6.45%
wlda	Test.1	11.74%	12.47%
	Test.2	5.86%	6.40%
Decision tree T2			
type of lda	testset	70K	40K
lda	Test.1	-	11.60%
	Test.2	-	5.45%
wlda	Test.1	-	12.00%
	Test.2	-	5.53%

Table 1: Baseline systems

3. ERROR ANALYSIS OF THE UNSUCCESSFUL CALLS

In this section we present a full analysis of the errors that include both speech recognition related and non speech recognition errors.

The categorization of the non speech recognition errors is shown in Table 2. The analysis is made from 380 unsuccessful calls.

Some names such as foreign names are hard to pronounce for speakers who are not familiar with the names and that there are no standardized methods for pronouncing proper names. This causes callers to not complete the pronunciation of the name within the specified time duration. This also leads to pauses and hesitations.

Type of errors	# of calls	% of calls
Words that are not names	17	4.5%
Last name only	48	12.6%
First name only	34	8.9%
Wrong version of first name	40	10.5%
Duplicated first name	7	1.8%
First/last name reversed	5	1.3%
Names that are hard to pronounce	75	19.8%
OOV names	140	36.8%

Table 2: Categorization of non speech recognition errors

Many of the first names have shortened versions, or nicknames. In our vocabulary and finite-state grammars, it is not possible to include all such versions or combinations of the same. If a caller uses a nickname which is not in the vocabulary or the grammar, or a nickname followed by the formal first name and the last name, the call will fail. Table 3 presents a few such examples.

In Table 2, a large portion of unsuccessful calls is due to OOV names (36.8%). This indicates that we need a good confidence measure scheme to reject such calls or to remind callers to verify the names they are calling in order to increase the system usability.

Homonyms are another source of error. In the situation that "SMITH", "SMYTH" or "SMYTHE" is a last name, the first name will decide the person being called. However if the first name is also ambiguous, for example, "VICKI", "VICKY", or "VICKIE", this becomes a more difficult problem to solve.

In vocabulary	Caller said
Robert Dalrymple	Rob Dalrymple
Robert Dalrymple	Rob Robert Dalrymple
Eugent Clark	Gene Clark
Andrew Waldrow	Andy Waldrow

Table 3: In-consistency in first names

Table 4 presents the break down of speech recognition related errors from the baseline system. Some errors are combinations of different types as indicated in the second part of the table. The major sources of speech recognition error appear to be noise related errors, normal speech recognition errors and pronunciation errors. Section 5 addresses the first two types of errors and Section 6 addresses the pronunciation errors.

Type of errors	% of calls
Noise related errors	30.6%
Normal speech recog. errors	34.6%
Pronunciation errors	21.5%
Spelled out names and chopped speech	13.2%
The overlap between categories	
Pron. errors and noise	5%
Pron. errors and chopped speech	9.5%
Noise and chopped speech	3.5%

Table 4: Categorization of speech recognition errors

4. EFFECT OF THE LEXICON AND FINITE-STATE-GRAMMAR SIZES

A significant challenge of the task is its large vocabulary. As expected, an increase in the size of the vocabulary and the grammar (allowed legal names) results in increased error rates. See results in Table 5. The system (lda, 70K Gaussian) trained off Decision tree T1 was used in this experiment.

However, an increased number of pronunciation variants alone, i.e., increasing the size of lexicon, without an increase in the grammar size does not add to the acoustic confusability. See third column in Table 5. In fact, this leads to lower error rates. This is primarily due to the fact that names are pronounced very differently by different people and the more pronunciation variations we can capture, the better the performance.

	Grammars and lexicons		
testset	G1, L1	G2, L2	G1, L2
Test.1	11.77%	22.7%	-
Test.2	5.75%	13.8%	5.25%

Table 5: Effect of grammar and lexicon sizes

5. CLUSTERING AND ADAPTATION

Speaker clustering has been shown to be effective to improve speech recognition accuracy in large vocabulary, dictation tasks [7]. A second reason for clustering to be helpful in the name recognition task is that by clustering the training data, different channels and noise (calling) conditions can be modeled by different cluster models. During recognition, a cluster that best models the test speaker and the channel condition is selected. On the other hand, speaker adaptation is an effective way to bring the performance of a speaker independent system to be closer to the performance of a speaker dependent system. In this section, we present variations of speaker clustering and adaptation methods that provided significant gains on the large vocabulary name recognition task.

5.1. Speaker Clustering

We built speaker clustered systems with 2 and 8 clusters. Simple HMM models that have one Gaussian per context independent phone state were obtained for each speaker from the training data. Then, the means were clustered using the k-means algorithm [7]. This was done for speakers that have more than 50 utterances in the training data.

The optimally selected cluster is the one that yields the maximum likelihood for each test utterance. Table 6 shows that a 12.7% relative improvement can be obtained from the 8-cluster models. In this table, WER refers to the word error rate and SER refers to the sentence error rate.

When the test utterances are very short, we believe that the best way of using clustered models is by interpolating between them using a cluster weighting scheme and combining with speaker adaptation to achieve better performance. The details are presented in Section 5.3.

Model	WER	SER
baseline	11.60%	12.37%
2-cluster	10.83%	11.97%
8-cluster	10.13%	11.39%

Table 6: Results of speaker clustering

5.2. Massive Adaptation

Classically in speech recognition applications such as dictation, adaptation is performed using some adaptation data collected from the test speaker. Subsequently, the Gaussian means and the variances of the speaker independent models are adapted to this speaker. However, in telephony applications, especially in name dialing, it is not always possible to gather a lot of data from a single speaker. However, we observe from our test data that usually a person calls the same set of individuals, or when the call is not successful, the caller tries the same name repeatedly. Instead of using a generic telephony speech recognition system, it is beneficial to perform adaptation on the most recent calls to enhance the performance. We call this new procedure “massive adaptation” since the adaptation data is obtained from a pool of calls rather than from a single speaker. The adaptation algorithm used is the combination of MLLR[11] and MAP[10].

As mentioned earlier, we collected our name dialing data into two test sets, Test.1 and Test.2. Although there was no overlap between them, we observed that they had common characteristics. They were both obtained from the same name dialer used in IBM. They had some common speakers (possibly calling the same person). So, we adapted the general telephony system to the name recognition domain using Test.2 as adaptation data to do massive adaptation on the speaker independent acoustic models. Then we tested the performance of Test.1 before and after performing massive adaptation. The recognition accuracy improves significantly after massive adaptation, as shown in the second row of Table 7.

5.3. Unsupervised Utterance Adaptation

To improve the decoding accuracy from an unknown speaker, we can use the testing utterance itself to do the adaptation. This process must be unsupervised, since in reality, correct script is not available. The adaptation needs to be robust to avoid over training when the adaptation utterance is very short. This robustness can be achieved by using an adaptation algorithm that requires fewer parameters to be estimated, or using prior information to constrain the estimation.

A two-pass decoding is needed for each call. In the first pass, a speaker independent system or the system after massive adaptation is used to obtain a decoded script, then a forward-backward algorithm is performed on the decoded script to obtain adaptation statistics. After adapting the acoustic models using these statistics, a second pass decoding is performed using the adapted models. For details of the adaptation procedure, see [9].

We tried many different adaptation methods for our tests. Some results are shown below. Although full MLLR adaptation did not improve the error rate much, we had considerable improvements by doing block diagonal MLLR, cluster weighting and MAPLR adaptation. The best result is obtained from a modified version of MAPLR with a 12.6% relative error reduction. Details of the algorithms and experiments are provided in [9]. In Table 7, the models after massive adaptation are used as the baseline models. The adaptation statistics for utterance-based unsupervised adaptation, as shown in the last four rows on Table 7, are derived from the decoded transcriptions obtained using these models. For the Cluster Weighting and Bias method (CWB), the interpolating weights and the biases are estimated jointly using the clustered models (as in Section 5.1) and the statistics obtained from massive adaptation models. For the MAPLR scheme, the clustered models and the weights are used to estimate the priors for the transformations.

Method	WER	SER
baseline (SI)	11.60%	12.37%
baseline (massive adapted)	9.53%	10.52%
MLLR (full matrix)	9.39%	10.37%
MLLR (block diagonal)	8.83%	9.90%
CWB	8.79%	9.81%
MAPLR with CW prior	8.33%	9.27%

Table 7: Results of unsupervised adaptation

6. PRONUNCIATION MODELING

A careful analysis of the speech recognition errors pointed to the fact that about 21.5% of the recognition errors were accent or pronunciation related. Since names in general are more difficult to pronounce than other commonly used words, many of these errors are related to lack of knowledge on how to pronounce them and the differences that arise from native and non-native pronunciations. As a preliminary experiment to study the effect of automatically derived pronunciations from acoustic evidence alone on this name dialing task, the algorithm presented in [8] is used.

A trellis of sub phone units is constructed from the speech utterance. The probability of a transition occurring from one node to another in the trellis is determined by weighting the score obtained from a Hidden Markov Model (HMM) [8] with a precomputed node-to-node transition probability obtained from a database of names.

In our experiments, the algorithm was applied to sample utterances of every name in a subset of the test set. The derived pronunciation was then added to the lexicon and the remaining utterances in the test set were decoded using the new lexicon. A reduction in the error rate from 13.37% to 11.09%, i.e., relative 17.0% improvement, is seen if the acoustically derived pronunciations are added to the lexicon. When the same lexicon was used to decode a subset of Test.2, the improvement in error rate was not significant. This is explained by the fact that the names in Test.2 and Test.1 do not significantly overlap. Therefore, the newly added pronunciations were hardly used. However, careful analysis of the decoding results indicates that if an automatically derived pronunciation exists in the lexicon for a name in the test set, the decoder preferred this pronunciation to the linguistically generated one.

This algorithm can also serve as a means to derive personalized vocabularies. This feature will enable the user to add words to their personalized vocabulary, for which an a priori spelling or acoustic representation does not exist in the lexicon, and associate that word(s) to a phone number to be dialed. Once the personalized vocabulary is configured, the user can subsequently dial the phone number by speaking the new word(s) just added to the vocabulary.

Possible extensions to this kind of pronunciation modeling is discussed in Section 7.

7. CONCLUSIONS AND FUTURE WORK

Although this large vocabulary name recognition task is different from any other LVCSR task defined in the literature, algorithms used in LVCSR tasks are applicable to this task also. More specifically, rapid adaptation and unsupervised utterance adaptation techniques presented in this paper are extremely valuable to this kind of an application. This is primarily due to the small amount of data (3 seconds or less) available for adaptation. Speaker clustering and massive adaptation algorithms serve to match the test data with the training data, including channel and environment noise. The adaptation algorithms that have been described in section 5 have been very effective for this task. It is also important to model the phonetic contexts with clean data to eliminate any noisy alignments. The speaker independent models can then be built out of both clean and noisy data. Collectively, we have obtained gains in recognition accuracy of about 28% relative.

Inspired by the results of our experiments in modeling pronunciation variations, a syllable-based approach where the syllables are trained specifically on the data from names is currently being explored. In the future, we plan to focus on robustness to noise from car, cellular telephones, etc.. Incorporation of nicknames and repetitive usage of different variations of first names into the grammar will increase the success rate of the calls. Use of spelled names for improving recognition when the confidence measure associated with the recognized utterance falls below a threshold, is an area that requires further investigation. Confidence measures can also be used to reject OOV names or to verify the recognition results.

Making the name dialing application more conversational will increase the user friendliness and also reduce the non speech recognition related errors. Introduction of hierarchical grammars, that are organized based on location of the individuals being called will resolve some amount of ambiguity and help in increasing recognition accuracy.

8. REFERENCES

[1] C.S. Ramalingam, L.P. Netsch, and Yu-Hung Kao, “Speaker independent name dialing with out-of-vocabulary rejection,” Proceedings of IEEE ICASSP’97, Vol. 2, pp 1475-1478.

[2] J.M. Elvira and J.C. Torrecilla, “Name dialing using final user defined vocabularies in mobile (GSM and TACS) and fixed telephone networks,” Proceedings of IEEE ICASSP’98, Vol. 2, pp 849-852.

[3] C.S. Ramalingam, Yifan Gong, L.P. Netsch, W.W. Anderson, J.J. Godfrey, and Yu-Hung Kao, “Speaker dependent name dialing in a car environment with out-of-vocabulary rejection,” Proceedings of IEEE ICASSP’99, Vol. 1, pp 165-168.

[4] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, M. A. Picheny, “Robust Methods for Using Context-Dependent Features and Models in a Continuous Speech”, Proceedings of IEEE ICASSP’94, Vol. 1, pp 533-536.

[5] Y. Li, Y. Gao, H. Erdogan, “Weighted Pairwise Scatter to Improve Linear Discriminant Analysis”, Proceedings of ICSLP2000, Vol. 4, pp 608-611.

[6] Random House Webster’s Unabridged Dictionary, Random House, New York, 1998

[7] Y. Gao, M. Padmanabhan, M. Picheny, “Speaker Adaptation Based on Pre-Clustering Training Speakers”, Proceedings of EuroSpeech’1997, pp 2091-2094.

[8] B. Ramabhadran, L. R. Bahl, P.V. deSouza, M. Padmanabhan, “Acoustics-Only Based Automatic Phonetic Baseform Generation ”, Proceedings of ICASSP 1998, Vol. 1, pp:309 -312, 1998.

[9] H. Erdoğan, Y. Gao, M. Picheny. “Rapid Adaptation Using Penalized-Likelihood Methods”, submitted to ICASSP2001.

[10] J. L. Gauvain and C. H. Lee, “Maximum-a-Posteriori estimation for multivariate Gaussian observations and Markov chains”, IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp 291-298, April 1994.

[11] C. J. Leggetter, et al, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, Computer Speech and Language, Vol 9, 1995.