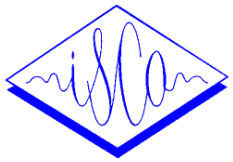

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Towards Robust Telephony Speech Recognition in Office and Automobile Environments

Subrata Das, David Lubensky

IBM Thomas J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

ABSTRACT

This study is concerned with improving the robustness of our telephony speech recognition system. Our previous implementation of this system handled both landline and cellular speech produced in a relatively quiet environment, such as in a regular office. However, it was found to be unduly vulnerable to background noise. In particular, we wanted to improve the accuracy of the system in the environment of a moving automobile, without impacting its performance in a quieter background. We collected samples of automobile noise under various operating conditions. We used these noise files to artificially generate noisy training data. We applied MAP adaptation procedure to study the value of such noisy training data and found that it helped to improve the robustness of our telephony speech recognition system. In a related study, we also investigated performance scores as a function of the total number of vocabulary entries in the system, as this has implications for a practical system implementation.

1. Introduction

We find ourselves in a world where computers for information exchange are destined to be pervasive and ubiquitous. Telephony speech recognition promises to play a central role in this information revolution. It will empower us to carry out transactions such as checking our e-mails, buying and selling stocks and making a travel reservation, from anywhere and at anytime.

Our previous study in telephony speech recognition was reported in the Proceedings of the Eurospeech '99 [1]. The training data used to construct this system consisted of both landline and cellular data collected in typical office environments. When tested on a list of stock names and digit strings, it performed relatively well in a quiet environment. However, it was found to be vulnerable to background noise, such as that encountered in an automobile. Consequently, the intention of our current study was to investigate strategies for enhancing the robustness of our system under such ambient noise exposure, without sacrificing the recognition accuracy in a quiet environment. For practical reasons, we wanted to develop a single versatile system capable of handling the multitude of channel and environmental combinations. Studies on speech recognition in a car environment were reported previously [2, 3]. However, these studies were concerned with on-board applications and did not utilize cellular data.

The paper is organized as follows. In Section 2, we describe the training and test databases that we utilized to carry out our experiments. In addition to landline and cellular speech data collected from a large number of speakers, the training database included recordings of samples of noise typically present in stationary and moving automobiles. The noise data helped to improve the robustness of our system. The test database consisting of speech collected from a fixed group of speakers in a variety of environments, both stationary and mobile, was intended to permit meaningful comparison of system performances under these environments. In section 3, we introduce the framework of our speech recognition system and proceed to discuss the experimental studies. An associated aspect of our study dealt with the change in system performance as a function of the vocabulary size. We also examined cellular artifacts such as fading and dropouts and observed how they tended to influence the performance of our recognition system. Finally, in section 4, we summarize the results of our studies and com-

ment on some possible avenues for further work in this area.

2. Databases

We describe our training and test databases in this section. The speakers contributing training data read from prepared texts drawn from several categories which are deemed to be useful in our vision of an ubiquitous computing world of the future. Some examples of these categories are common first and last names of people, digit strings and stocks and mutual fund names.

Our training databases were of three distinct types. Two of these were described earlier [1], but will be summarized here for the sake of completeness. The first one consisted of landline recordings of 255,000 sentences from 25,000 speakers. The second type was comprised of recordings made over all three cellular channels, CDMA, TDMA and GSM, prevalent in the United States, using hand-held cellular phones. Our cellular data, which were collected in ordinary office environments, totaled 76,000 sentences spoken by 760 speakers.

The third type of training data were not of speech, but consisted of recordings of noise typically present in stationary and moving automobiles. As explained in the next section, we utilized this noise database to study the value of artificially generating noisy training data by judiciously mixing noise to our relatively quiet cellular speech database. The noise recordings were made over all three cellular channels under a variety of common operating conditions in a car. For instance, we collected some noise data with the air-conditioner fan running and some with the fan turned off. Similarly, other factors that determine noise characteristics, such as window opening and car speed were included. Sum of all the noise data amounted to approximately 30 minutes of recording. These three types of training data are listed in Table 1.

Channel	No. of speakers	No. of sentences
Landline	25000	255000
Cellular	760	76000
Cellular	Automobile noise	30 minutes

Table 1: Training databases

Our test database was tailored to permit, within limits of practicality, a reasonable comparison of our system performance under differing ambient and channel conditions. Cellular data encompassed CDMA, TDMA and GSM channels. All landline recordings were done in office environments. Cellular recordings took place in an office as well as in a car driven at 30 and 60 miles per hour. The same group of 20 male and 20 female speakers, who never contributed any training data, participated in all of the 10 test data recording sessions. These are listed as session numbers 1 through 10 in Table 2. The rows marked A00, A30 and A60 are used to tabulate the average results for cellular channels in an office environment, in a car at 30 mph and in a car at 60 mph respectively. All test data were collected using hand-held phones and a fixed list of stock names.

3. Experimental Studies

In this section, we describe our experimental work to improve the robustness of our telephony speech recognition system and carry out other related studies. Our general approach was to develop a new recognition system by applying MAP adaptation technique [4] on the IBM Conversational Telephony System for Financial Applications (IBM-CTSFA) [5]. The IBM-CTSFA was generated earlier by employing only the landline training database listed in Table 1. It was based on our conventional steps for MFCC signal processing, subphonetic labeling using decision tree networks, a fast match stage to trim the search list and a detailed match procedure for final selection in association with a stack decoder.

Our first task was to establish a performance

baseline, so that subsequent results could be compared against this baseline. The baseline system was designed using the landline and the cellular training databases listed in Table 1, disregarding the third training database of automobile noise. Signal processing consisted of deriving a set of parameters representing the 12-th order Mel frequency cepstral coefficients [6] and an energy estimate for every 10 msec of time frame. This was supplemented by first and second order derivatives of these parameters as a function of time. Application of MAP adaptation procedure with these data on the IBM-CTSFA generated the baseline system. We tested our baseline system on the 10 sets of test data, listed in Table 2, which were the recordings of some stock names. Our entire list of stock names consisted of 23,000 entries. We utilized a finite state grammar with equal probability for each of these stock names during decoding. The corresponding word error rates are tabulated under the "Base" column in Table 2. We see that the baseline system performed relatively well in an office environment. For instance, average word error rate (WER) over the cellular channels was 10.7 percent. However, in an automobile moving at 30 miles per hour, the performance fell to 16.7 percent. When the speed increased to 60 miles per hour, performance further degraded to 29.7 percent WER on the average. Both increased levels of noise and cellular artifacts, such as fading and dropouts, were responsible for the performance deterioration at higher speeds. For instance, we estimated that almost 57.1 percent of the errors in the GSM test data recorded at 60 mph speed were related to fading and dropout problems.

Our next experiment utilized all of the three training datasets. We artificially generated noisy cellular training data as follows. We took contiguous segments of noise starting at random times from the third training dataset of automobile noise and mixed them with the files of our celular speech training data. After some preliminary experimentation, we decided to create two sets of such noisy training data, one to roughly match the background level of noise at 30-mph and the other to more or less correspond to 60-mph noise level. Next, the four training datasets – the landline data, the (original) cellular data, the (30-mph) noise-mixed cellular data and the (60-mph) noise-mixed cellular data – were all applied for MAP adaptation. The WER for this system are listed in Table 2 under the column marked "N-mix". Comparing the "Base" and "N-mix" columns, we note that the average WER over cellular channels in the relatively quiet office environment was unchanged at 10.7 percent, although the landline performance degraded from 8.0 to 8.3 percent. The WER of the 30-mph test data was reduced from the baseline result of 16.7 percent to 15.5 percent. Similarly the WER on the 60-mph test data was 24.3 with the "N-mix" system, an improvement over the 29.7 percent WER observed for the baseline system.

No.	Location	Channel	Base	N-mix	voc/4	voc/8
1	Office	Landline	8.0	8.3	6.9	4.5
2	Office	CDMA	12.0	11.2	8.7	6.9
3	Office	TDMA	9.6	10.2	7.8	5.3
4	Office	GSM	10.5	10.7	8.4	6.3
A00	Office	Cellular	10.7	10.7	8.3	6.2
5	Car-30	CDMA	16.9	15.8	13.1	9.9
6	Car-30	TDMA	15.7	15.2	10.9	7.5
7	Car-30	GSM	17.3	15.5	12.7	9.0
A30	Car-30	Cellular	16.7	15.5	12.2	8.8
8	Car-60	CDMA	29.8	23.1	16.9	12.4
9	Car-60	TDMA	30.3	25.4	20.4	17.1
10	Car-60	GSM	29.2	24.4	19.9	16.3
A60	Car-60	Cellular	29.7	24.3	19.1	15.3

Table 2: Test sessions 1 through 10 with a comparison of word error rates. Rows A00, A30 and A60 list the average error rates for cellular data as shown.

We mentioned that our list of stock names had 23,000 entries. In practice, if the accuracy obtained with such a large list is unacceptable, it is possible to introduce an additional processing stage to divide up this single large list into a number of lists, each with a smaller number of entries. For instance, the list may be divided into four equal parts, each with a quarter of the

entries, and the user required to choose one of the four parts first, before asking for the particular stock of interest. With this type of application in mind, we studied the system performance as a function of the total number of entries. Columns marked "voc/4" and "voc/8" of Table 2 show the results of this experiment when the the number of entries were reduced to a quarter and an eighth of the original respectively. As expected, the performance improved with less number of entries. For example, the average error rate for the automobile data collected at 30 miles per hour went down from 15.5 percent to 8.8 per cent when the number of entries was reduced from 23,000 to approximately 2900 stock names.

4. Conclusions

Our previous telephony speech recognition system [1], capable of handling both landline and cellular data, was found to be unduely sensitive to background noise in an automobile. We were able to improve our system performance by collecting samples of automobile noise, mixing this noise with our relatively quiet cellular training data and carrying out MAP adaptation. We realized performance gains at 30 and 60 miles per hour speeds with little degradation to the office environment accuracy. In addition to the increased background noise level, cellular artifacts such as dropouts and fading influenced the system performance at higher speeds.

We also studied how the recognition performance changed as the system perplexity was reduced by limiting the number of entries in the system. These results would be helpful in designing a system for practical use.

While it was possible to somewhat compensate for increased ambient noise at higher automobile speeds, as shown in this paper, cellular artifacts such as fading and dropouts call for a different strategy. Segments of speech ranging from several centiseconds to a second or so may be missing in some cases due to these artifacts. We plan to employ a type of confidence measure [7] to handle such cases, so that, if needed, the user can be prompted for another attempt. We also intend to deal with the issues related to far-field effects in speech recognition which may be important for hands-free operation in a car.

5. References

1. S. Das, D. Lubensky and C. Wu, "Towards robust speech recognition in the telephony network environment - cellular and landline conditions," Proceedings of the Eurospeech '99, pp. 1959-1962, September 1999.

2. R.Bippus, A.Fischer and V. Stahl, "Domain adaptation for robust automatic speech recognition in car environments," Proceedings of the Eurospeech '99, pp. 1943-1946, September 1999.

3. M. Westphal and A. Waibel, "Towards spontaneous speech recognition for on-board car navigation and information systems," Proceedings of the Eurospeech '99, pp. 1955-1958, September 1999.

4. L. Bahl, F. Jelinek and R. Mercer, "A maximum likelihood approach to continuous speech recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 179 - 190, March 1983.

5. K. Davies, et al., "The IBM conversational telephony system for financial applications," Proceedings of the Eurospeech-99, pp. 2755-278, September 1999.

6. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, pp. 357-366, August 1980.

7. Q. Lin, D. Lubensky and S. Roukos, "Use of recursive mumble models for confidence measuring," Proceedings of Eurospeech-99, pp. 53-56, September 1999.