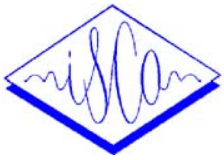

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# TOWARDS ROBUST SPEECH RECOGNITION IN THE TELEPHONY NETWORK ENVIRONMENT - CELLULAR AND LANDLINE CONDITIONS

*Subrata Das, David Lubensky, Cheng Wu*

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

10.21437/Eurospeech.1999-489

## ABSTRACT

We describe several speaker-independent speech recognition studies conducted with both landline and cellular network telephone data. The cellular environment included the three dominant standards found in the United States: CDMA, TDMA and GSM. Our goal was to design a system that operated over all these four channels, handling their innate variations, such as those of background and line characteristics. Our baseline system was trained on 200 hours of landline telephone speech from about 25,000 speakers. We experimented with MAP adaptation utilizing some training sentences from our cellular and landline databases. We applied an LDA procedure to improve our performance. We compared the performances of these systems on several independent test databases and demonstrated the effectiveness of a hybrid system built with data taken from all four networks.

## 1. INTRODUCTION

With voice-enabled applications such as stock quotations, call center operations and e-mail transactions on the horizon, speech recognition in the telephony environment is poised to take the center stage. Concurrently, the telephony environment itself is going through a series of rapid changes. Traditional landline phones are supplemented and on occasions outright replaced by the cellular ones. Some countries with a dearth of landline infrastructure have decided to switch directly to the cellular networks. Analog cellular channels are giving way to digital ones due to their superiority in a number of areas, such as bandwidth efficiency, transmission security and low power requirements. Under these contexts, our goal was to utilize our landline and cellular databases to investigate the performance of a standard telephony speech recognizer and then seek ways to improve its operation by applying signal processing and acoustic modeling techniques.

Cellular service providers in the United States generally subscribe to one of three digital standards: CDMA (code-dependent multiple access), TDMA (time-dependent multiple access) and GSM (global system for mobile). Advocates of each system point to its own advantages. For instance, GSM phones can operate both in Europe and in the United States. CDMA with its spread spectrum technology is considered by many to be more secure and efficient. TDMA which encompasses the GSM technology is probably the most widely used system at this time.

Several researchers have suggested different techniques to enhance speech recognition performance in a cellular environment. For instance, Dufour, Glorion and Lockwood [1] developed a root-normalized front-end to handle GSM network data.

Our goal was to design a single system for all four networks: landline, CDMA, TDMA and GSM. To this end, we collected speech data through each of these three digital cellular networks as well as through regular landlines. The speakers read scripts prepared from a number of lists such as stock and mutual fund names. We describe the details of these databases in Section 2.

In Section 3, we describe our experimental work. Our baseline system [2], constructed exclusively from 200 hours of landline telephone data, worked

Table 1. Training and test databases

| Data | Number of Speakers | Number of Sentences | Script type |
|---|---|---|---|
| Landline Training | 25,000 | 255,000 | Mixed |
| Cellular Training | 760 | 76,000 | Mixed |
| Landline Test | | 2,000 | Stock names |
| CDMA Test | 40 | 802 | Stock names |
| TDMA Test | 40 | 759 | Stock names |
| GSM Test | 40 | 806 | Stock names |
| Dig7 Test | 278 | 323 | Digits |
| Dig10 Test | 686 | 1150 | Digits |

well on landline test data, but relatively poorly on cellular network data. Starting from this baseline system, we employed MAP adaptation [3] to build other recognition systems. Another aspect of our work was concerned with our studies with the LDA (linear discriminant analysis) technique [4, 5] to boost the performances of our systems. We evaluated and compared the performances of these systems utilizing a number of test databases.

We conclude this paper with a summary of our work and some relevant observations in Section 4.

## 2. DATABASES

As mentioned in Section 1, we experimented with a number of databases to construct and test our recognition systems. The speakers always read from prepared texts drawn from several categories, such as, stock and mutual fund names, first and last names of people and digit strings. These databases are listed in Table 1.

The "Cellular Training" database is a mixture of speech collected through all the three cellular media. For practical reasons, we decided to pool together all cellular data for building our system, rather than construct a separate system for each type. In all the databases, the speakers were selected more or less equally from both genders. Our primary test datasets, Landline Test, CDMA Test, TDMA Test and GSM Test, consisted of utterances of stock names. But each dataset was comprised of speech from a different set of speakers reading a different set of scripts. Consequently, the error rates noticed for different cellular channels were not rigorously comparable to each other. However, we believe our test databases were sufficiently large to permit a reasonable evaluation.

We used the test datasets Dig7 Test and Dig10 Test for some additional testing. These were 7- and 10-digit long strings of utterances, continuously spoken, which were recorded over regular landlines.

## 3. EXPERIMENTAL STUDIES

We constructed four recognition systems with the databases described in the previous section. Signal processing for the first three systems consisted of deriving a 12-th order MFCC (Mel frequency cepstral coeffecients) [6] and energy every 10 msec, along with their first and second derivatives. The fourth system resorted to an additional step of calculating a set of LDA parameters [4, 5] every 10 msec. The following is a summary of these four systems which were derived from our baseline system by using MAP adaptation technique.

1. Cell52: This was a cellular system constructed by utilizing 52,000 sentences of our cellular training data.

2. Cell76: This was another cellular system where we used all 76,000 cellular training sentences.

3. HybridM: This was the first hybrid system based on both cellular and landline data, using MFCC and derivatives for signal processing.

4. HybridL: This was the second hybrid system to employ both cellular and landline data, but in this case we obtained LDA parameters during signal processing.

Our primary testing was carried out with the test databases of 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 names during decoding. The first study showed the effect of training database size

Table 2. Comparison of word error rates (%) between two cellular recognition systems

| Test Data | Cell76 System | Cell52 System |
|---|---|---|
| CDMA Test | 7.9 | 8.4 |
| TDMA Test | 13.0 | 14.2 |
| GSM Test | 11.9 | 13.9 |
| Average | 10.9 | 12.2 |

on performance accuracy. Table 2 compares the word error rates of Cell52 and Cell76 systems on the cellular test databases. We saw that the system performance was significantly enhanced for all three cellular channels with the use of more training data. Word error rate decreased from 12.2 percent to 10.9 percent on the average. We also noticed that the CDMA data had significantly better recognition score than the other two cellular varieties. However, we mentioned in Section 2 that these three test datasets were not strictly comparable to each other.

Next, in Table 3 we compare the word error rates of the four systems: Cell76, Baseline, HybridM and HybridL. Paying attention to only the first three systems with identical signal processing procedures, we observed that cellular test data had their best scores on the cellular system Cell76. Similarly, the landline test data achieved the least error rate on the landline (baseline) system. The HybridM system was a good compromise, as its landline performance matched the baseline one and its cellular results were only moderately worse than those of the Cell76 system. The best overall performance was observed for the HybridL system, which utilized LDA parameters. Its error rates on cellular data were comparable to rates observed for the Cell76 system. In addition, its landline results were the best yet. Notice that the landline test data had poorer recognition score than the cellular ones. We attributed this behavior to the more noise observed in our landline test database.

Finally, as an added confirmation of the utility of

Table 3. Comparison of word error rates (%) between the four recognition systems

| Test Data | Cell76 System | Baseline System | HybridM System | HybridL System |
|---|---|---|---|---|
| CDMA Test | 7.9 | 13.7 | 9.6 | 7.9 |
| TDMA Test | 13.0 | 21.3 | 14.6 | 13.2 |
| GSM Test | 11.9 | 19.2 | 14.8 | 11.2 |
| Landline Test (Noisy) | 20.7 | 18.7 | 18.7 | 15.9 |

Table 4. Comparison of word error rates (%) on Dig7 and Dig10 databases

| Test Data | Baseline System | HybridL System |
|---|---|---|
| Dig7 | 1.6 | 1.7 |
| Dig10 | 1.3 | 1.3 |

the HybridL system, we tested the Dig7 and Dig10 datasets as well. These results are listed in Table 4. Once again, we see that the HybridL system compared favorably with the Landline system on Landline test data.

## 4. CONCLUSIONS

We conducted a number of experiments with the intention of developing a speech recognition system that can work well on both cellular and landline data. We used two procedures to achieve our goal: use of the LDA parameters for signal processing and use of the MAP adaptation strategy during the acoustic modeling stage. Our databases consisted of speech recorded via landline as well as via three major types of cellular networks. We conducted experiments to study several items of interest. Our first experiment demonstrated the importance of using a large database for training.

In the next experiment, we compared the performances of several systems. We concluded that a hybrid system based on LDA parameters and built with MAP adaptation of both landline and cellular data was our best system. We verified this by a further test on two databases of digit strings,

We plan to continue with further improvements to our system. For instance, we know that cellular channels are subject to miscellaneous vagaries peculiar to these media, such as fadeout, not found in landline data. Thus, in the case of fadeout, we would like to use a form of confidence measure [7] to decide if the total speech input is properly received and decoded, or, if the user needs to be prompted for another attempt. Finally, as we pointed out in Section 1, the whole field of telephony is in the middle of a flux of changes. Consequently, our future strategies need to be tailored to fit those changes.

## 5. ACKNOWLEDGEMENTS

Our baseline system was constructed by E E Jan and Mukund Padmanabhan using landline training data and MFCC signal processing.

## 6. REFERENCES

1. S. Dufour, C. Glorion and P. Lockwood, "Evaluation of the root-normalised front-end (RN-LFCC) for speech recognition in wireless GSM network environments," Proc. 1996 International Conference on Acoustics, Speech and Signal Processing, pp. 77 - 80, March 1996.

2. K. Davies, et al., "The IBM conversational telephony system for financial applications," Proceedings of Eurospeech-99, (this issue).

3. 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.

4. R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," Proc. 1992 International Conference on Acoustics, Speech and Signal Processing, Volume 1, pp. 13 - 16, March 1992.

5. R. Haeb-Umbach, D. Geller and H. Ney, "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," Proc. 1993 International Conference on Acoustics, Speech and Signal Processing, Volume 2, pp. 239 - 242, March 1993.

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, (this issue).