PERFORMANCE ANALYSIS OF THE AURORA LARGE VOCABULARY BASELINE SYSTEM[1]

N. Parihar and J. Picone
Inst.for Signal and Info. Proc.
Mississippi State University
{parihar,picone}@isip.msstate.edu

D. Pearce
Speech and MultiModal Gp.
Motorola Labs, U.K.
bdp003@motorola.com

H. G. Hirsch
Dept. of Elec. Eng. and Comp. Sc.
Niederrhein University
hans-guenter.hirsch@hs-niederrhein.de

ABSTRACT

In this paper, we present the design and analysis of the baseline recognition system used for ETSI Aurora large vocabulary (ALV) evaluation. The experimental paradigm is presented along with the results from a number of experiments designed to minimize the computational requirements for the system. The ALV baseline system achieved a WER of 14.0% on the standard 5K Wall Street Journal task, and required 4 xRT for training and 15 xRT for decoding (on an 800 MHz Pentium processor). It is shown that increasing the sampling frequency from 8 kHz to 16 kHz improves performance significantly only for the noisy test conditions. Utterance detection resulted in significant improvements only on the noisy conditions for the mismatched training conditions. Use of the DSR standard VQ-based compression algorithm did not result in a significant degradation. The model mismatch and microphone mismatch resulted in a relative increase in WER by 300% and 200%, respectively.

1. INTRODUCTION

Mobile computing devices lack sufficient computing resources to perform large vocabulary continuous speech recognition (LVCSR). Client/server architectures are one potential solution to this bottleneck. Mobile devices have sufficient computing resources to handle a few components of the recognition system, such as feature extraction. A popular architecture for such applications is the Client/Server Distributed Speech Recognition (DSR) architecture [1], shown in Figure 1. The main advantage of this approach is the ability to extract features on mobile terminal devices that can exploit sophisticated noise enhancement techniques to improve the overall recognition performance.

The goal of the ETSI ALV evaluation was to measure the relative performance of different front ends on a large vocabulary system using sub-word models to supplement the performance calibration on small vocabulary using word models [1]. A noisy version of the WSJ0 database was chosen as the large vocabulary task [2,3]. The baseline recognizer used for the ALV was developed by ISIP [4]. This paper presents design issues associated with the evaluation database and the baseline recognition system. An extensive analysis of the performance of the ETSI WI007 front end is also presented. Six focus test conditions were calibrated: *sampling frequency reduction*, u*tterance detection, feature vector compression*, *model mismatch*, *microphone variation*, and *additive noise*.

2. EXPERIMENTAL DESIGN

The 5K-word task for the WSJ0 Corpus was selected for the ALV evaluation because it represents a well-established LVCSR benchmark and constitutes a good trade-off between computational resources and complexity. The Nov'92 NIST evaluation set was used as the evaluation data set. Because the original WSJ data was collected at 16 kHz, an 8 kHz down sampled version was created. Processed versions of the data were created to simulate both filtered and additive noise conditions. A filtered version of the SI-84 training set for the Sennheiser microphone was used to construct the first training set, denoted as Training Set 1 (TS1).

For the second training set, the filtered SI-84 utterances were divided into two subsets: half recorded with the Sennheiser microphone and half recorded with a second microphone. No noise was added to one-fourth (893 utterances) of each of these subsets. To the remaining three-fourths (2,676 utterances) of each of these subsets, six different noise types (car, babble, restaurant, street, airport, and train) were added at randomly selected SNRs between 10 and 20 dB. The goal was an equal distribution

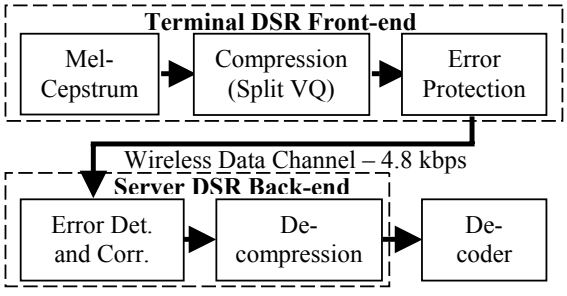

Figure 1: The Aurora standard for a DSR architecture.

1. This material is based upon work supported by the European Telecommunications Standards Institute (ETSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of ETSI.

of noise types and Signal to Noise Ratios. Thus, one clean set (893 utterances) and 6 noisy subsets (446 utterances each) were obtained for each of the two microphone conditions to systematically test the microphone and noise conditions. Each of the filtered versions of the evaluation set recorded with the Sennheiser microphone and second microphone. were selected to form two of the 14 evaluation sets (sets no. 1 and 7 respectively). The remaining 12 subsets were defined by adding each of the 6 noise types at randomly chosen Signal to Noise Ratios between 5 and 15 dB for each of the two microphone types.

All baseline experiments employed state-tied cross-word speaker-independent triphone acoustic models with four Gaussian mixtures per state. A single-pass Viterbi beam search-based decoder was used along with a standard 5K lexicon and bigram language model. The pronunciations in the lexicon were extracted from the publicly available CMU dictionary (v0.6) [5] with some local additions.

The baseline system used in the evaluation was modeled after a 16-mixture WSJ0 system with a WER of 8.3%. Table 1 shows a comparison of this system to the state-of-the-art for a variety of published systems [4]. It was decided that adaptation or proprietary lexicons would not be used in this evaluation, which accounts for a large part of the variation in performance shown in Table 1.

The ETSI WI007 front end [1] was chosen as baseline front end for the ALV evaluation. This front end is based on the standard mel frequency-scaled cepstral coefficients (MFCCs) and includes a lossy vector quantization compression algorithm that reduces the transmission bit rate to 4800 b/s.

There was a strong interest in reducing the computational requirements to conduct the evaluation. We followed a three-step approach to reduce the overall computation time without significantly compromising the quality of the evaluation:

- reduce the size of the test set by 50%;
- adjust beam pruning to reduce decoding time 6x;
- use only 4 mixtures per state.

The impact of these changes on performance, shown below in Table 2, is described extensively in [4].

3. ANALYSIS

The analysis of the ETSI front end on six focus conditions is described below. All experiments were analyzed using the MAPSSWE significance test with a significance level equal to 0.1%.

Site	Acoustic Model	Language Model	WER
ISIP	xwrd/gi	Bigram	8.3%
CU [6]	xwrd/gi	Bigram	6.9%
LT [7]	xwrd/gi	Bigram	6.8%
CU [6]	xwrd/gi	Bigram	6.6%
UT [8]	xwrd/gi	Bigram	6.4%

Table 1: A comparison of performance reported in the literature on theWSJ0 SI-84/Nov'92 evaluation task.

3.1. Sampling Frequency Reduction

First we explored the influence of sampling frequency reduction. For Training Set 1 (TS1), degradations due to a reduction in sampling frequency from 16 kHz to 8 kHz did not follow any trend. However, as shown in Figure 2, for Training Set 2 (TS2), significant degradations in performance were observed on the Sennheiser conditions (Test Sets 3-7). Statistically significant test conditions are indicated in bold.

The overall frequency response of the two microphone conditions is shown in Figure 3. The Sennheiser condition, as expected, preserves high frequency information better than the second microphone condition, resulting in slightly better performance at a 16 kHz sampling frequency. Surprisingly, a similar degradation due to sampling frequency reduction is not observed on matched conditions (training on TS1 and decoding on Test Set 1), which use the Sennheiser microphone. In this case, the additional information provided by high frequencies (between 4 kHz and 8 kHz) does not contribute to any additional improvement in performance. The spectral information provided by low frequencies (below 4 kHz) is sufficient to reach the upper bound on performance.

Factor	WER	Relative Degradation
Baseline system (ISIP)	8.3%	N/A
Terminal Filtering (ISIP)	8.4%	1%
ETSI front end	9.6%	14%
Beam adj. (15xRT)	11.8%	23%
Reduce 16 to 4 mixtures	14.1%	20%
50% reduction of eval set	14.9%	6%
Endpointing silences	14.0%	-6%

Table 2: Relative degradation in WER due to the three-step approach used to reduce computational requirement.

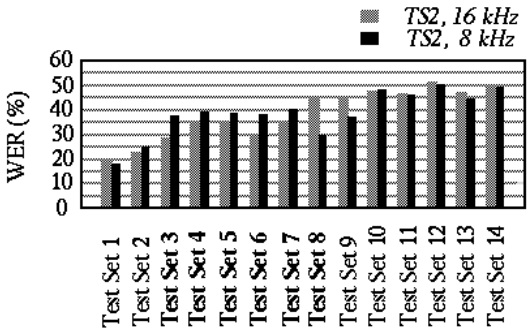

Figure 2: A comparison of the WER for 16 kHz and 8 kHz sample frequencies on TS2.

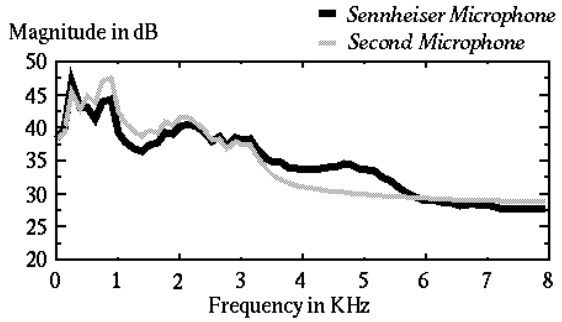

Figure 3: The Sennheiser close-talking microphone preserves frequencies above 3.5 kHz better on the average than the variety of microphones used on the second channel.

3.2. Utterance Detection

Utterance detection has been used in previous Aurora evaluations to decouple noise cancellation strategies from feature extraction during speech intervals. Utterance detection resulted in a significant improvement in performance on Test Sets 2-14 when the system was trained on TS1 (clean training set). Two sample test conditions in Table 3 show that the reduction in insertion errors is primarily responsible for the improvement in performance. In this case, the "silence" model is not a good match to the background noise for the noisy conditions because it hasn't been exposed to that noise during training. Without endpointing, the noisy silences were interpreted as speech data, resulting in a higher insertion error rate.

In contrast, for TS2, a significant improvement in performance was detected only for Test Set 8 (a reduction in the number of deletions was primarily responsible for this improvement). Because the training conditions contained ample samples of the noise conditions, the non-speech segments were modeled adequately by the silence model. Hence, the insertion error rate did not increase significantly on the noisy test conditions.

3.3. Compression

No significant degradation in performance due to split vector (VQ) compression was detected for TS1 for both sample frequencies. Because there is no significant degradation for Test Set 1, which is a matched condition, we might draw a conclusion that the split VQ algorithm will not degrade the performance of the system.

However, there was a significant degradation in performance for five noisy conditions (3, 8, 9, 10, 12) at 16 kHz sampling frequency and two noisy conditions (7, 11) at 8 kHz sampling frequency on TS2. We have not found a consistent explanation as to why these particular noise conditions were adversely affected, but we believe it warrants a closer study of the behavior of compression algorithm for noisy data.

Set	W/O Endpointing			With Endpointing		
	Sub.	Del.	Ins.	Sub.	Del.	Ins.
2	41.4	3.6	20.1	40.0	3.6	13.0
9	54.4	12.3	15.1	49.1	15.1	10.1

Table 3. The primary reason for a reduction in WER on TS1 for utterance detection is shown to be a result of a reduction in the insertion error rate. The results are shown as percentages.

3.4. Model Mismatch

The best performance was observed on matched condition (TS1 and Test Set 1), when all the utterances were recorded with a Sennheiser microphone, as shown in Figure 4. Because training is based on a maximum likelihood parameter estimation process, high performance can only be achieved when the test conditions to generate feature vectors are similar to training conditions in terms of means, variances, etc.

For all other conditions involving TS1, the recognition performance degraded significantly. Because there are consistent differences in SNR, background noise, or microphone between the training and testing conditions, there were significant degradations in performance. Adaptation schemes might have remedied this problem. Systems trained on TS2 performed significantly better than those trained on TS1 across all noise conditions. These trends were consistent for both sample frequencies and both compression conditions.

3.5. Microphone Variation

In general, the Sennheiser microphone performed significantly better than the second microphone condition, as shown in Table 4. The first cell in this table corresponds to TS1, which consists of clean utterances recorded with a Sennheiser microphone, and Test Set 1, which consists of similar data. The second cell in the first row represents a mismatched condition in which the test set was recorded on a different microphone. There was a significant increase in WER, from 16.2% to 37.4%. The same argument of model mismatch discussed in the previous section can be extended to explain this degradation. The same trend is observed on the car noise condition (Test Sets 2 and 9).

TS2 has half of the utterances recorded on the same microphone and the other half on any one of the 18 microphone types. With Baum-Welch training, a maximum likelihood based parameter estimation method, models trained on TS2 quickly converge towards the Sennheiser microphone in terms of their means and covariances. Hence, both the clean (Test Set 1) and car (Test Set 8) conditions for the second microphone result in significant degradation in performance, as shown in the second row of Table 4. Also note that the last three cells in the second row, which correspond to various noise conditions, show less degradation in performance than the corresponding conditions in the first row. There is obviously value in

Training Set	Set 1 (Senn. Mic.)	Set 8 (Sec. Mic.)	Set 2 (Senn. Mic.)	Set 9 (Sec. Mic.)
1	16.2%	37.4%	49.6%	59.7%
2	18.4%	29.7%	24.9%	37.3%

Table 4. On TS1, performance drops due to a mismatch in microphones for the second microphone conditions. Performance on TS2 is slightly better for the noise conditions.

exposing the models to noise during the training process.

3.6. Additive Noise

Severe degradation is observed for all noise conditions and at both sample frequencies because no noise compensation or adaptive techniques were used for these evaluations. However, the severity of this degradation can be limited by exposing the models to noise conditions during the training process. In Figure 4 and Figure 5, we demonstrate that training the models on TS2, which contains samples of the noise conditions, reduces the severity of the degradation in the noisy conditions. A boldface label indicates statistically significant test conditions at a 0.1% significance level. An important point to note is that these degradations are still significant compared to the clean condition. Similar trends were observed when the feature vectors were compressed [4].

On TS1 and TS2, it is observed that performance on car noise conditions (Test Set 2) is better than for the other noise conditions (street traffic, train stations, babble, restaurants and airports). Because the car noise condition can be approximated as stationary noise, and the other noise conditions are heavily non-stationary, performance is significantly better. The simple silence model used can adapt to the background noise.

4. SUMMARY

In this paper, we have presented an LVCSR system that was developed for the ALV evaluation. This public domain system is based on the 5K WSJ0 task and achieved a performance of 14.0% WER. It runs at 4 xRT for training and 15 xRT for decoding on an 800 MHz Intel Pentium processor.

We also presented an analysis of the results from the baseline front end experiments. It is shown that increasing the *sampling frequency* from 8 kHz to 16 kHz resulted in significant performance improvement only for the noisy test conditions. *Utterance detection* resulted in significant improvements only on the noisy conditions for the mismatched training conditions. The DSR standard VQ-based *compression* algorithm did not result in a significant degradation in performance. A mismatch between training and testing conditions (*model mismatch*) resulted in a 300% relative increase in WER whereas the mismatches in microphones resulted in a 200% relative increase in WER.

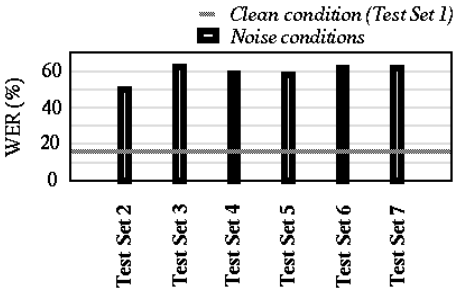

Figure 4: A comparison of the WER for six noise conditions at 8 kHz on TS1.

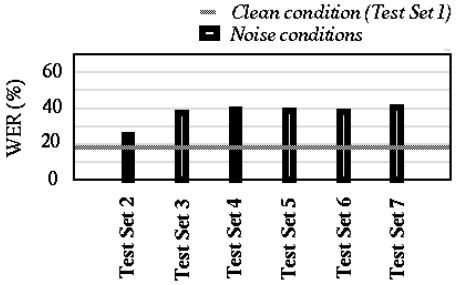

Figure 5: A comparison of the WER for six noise conditions at 8 kHz on TS2.

5. REFERENCES

1. "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front end Feature Extraction Algorithm; Compression Algorithm," *ETSI*, April 2000.

2. D. Pearce, "Overview of Evaluation Criteria for Advanced Distributed Speech Recognition," ETSI STQ-Aurora DSR Working Group, October 16, 2001.

3. G. Hirsch, "Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task," *ETSI STQ Aurora DSR Working Group,* June 2001.

4. N. Parihar and J. Picone, "DSR Front End LVCSR Evaluation," AU/384/02, Aurora Working Group, Dec. 2002.

5. "The CMU Pronouncing Dictionary," Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 2001.

6. P. C. Woodland, *et. al.*, "Large Vocabulary Continuous Speech Recognition using HTK," *Proc. of ICASSP*, Adelaide, Australia, pp. II/125-II/128, April 1994.

7. W. Reichl, and W. Chou, "Decision tree state tying based on segmental clustering for acoustic modeling," *Proc. of ICASSP,* pp. 801-804, Seattle, WA, USA, April 1998.

8. L. Welling, S. Kanthak, and H. Ney, "Improved Methods for Vocal Tract Normalization," *Proc. of ICASSP,* pp. 761-764, Phoenix, Arizona, USA, March 1999.