Performance Analysis of the Aurora Large Vocabulary Baseline System[1]

N. Parihar, and J. Picone *D. Pearce* *H. G. Hirsch*

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

Speech and Multimodal Group
Motorola Labs, U.K.
bdp003@motorola.com

Dept. of Elec. Eng. and Comp. Sc.
Niederrhein University (of App. Sc.)
hans-guenter.hirsch@hs-niederrhein.de

Abstract

In this paper, we present the design and analysis of a speech recognition system that was used to conduct the ETSI Aurora large vocabulary evaluation. The experimental paradigm is presented along with the results from a number of experiments designed to minimize the computational requirements for the system. 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 lossy 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. Mismatches in microphones also resulted in 200% relative increase in WER. The Aurora LV 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).

1. Introduction

Mobile computing devices still lack sufficient computing power and memory to perform large vocabulary continuous speech recognition (LVCSR). Client/server architectures are one potential solution to this bottleneck. Mobile devices do have sufficient computing resources to handle some components of the problem, such as feature extraction. One 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 small terminal devices which can exploit sophisticated noise enhancement techniques specific to the terminal device to improve recognition performance.

The goal of the ETSI Aurora large vocabulary (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 [2]. A noisy version of the WSJ0 database was chosen as the large vocabulary task [3,4]. The recognizer developed for these evaluations was based on the system developed by ISIP [5]. 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 [1] is also presented. Six focus test conditions were calibrated using speech recognition word error rate as the metric: *sampling frequency reduction* (16 kHz and 8 kHz), u*tterance detection* (influence of endpointing), *compression* (a vector quantization-based compression scheme), *model mismatch (*mismatched training and testing conditions), *microphone variation* (two microphone conditions available in the WSJ0 task [3]), and *additive noise (*six noise types collected from street traffic, train stations, cars, babble, restaurants and airports at varying signal-to-noise ratios).

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

2. Experimental Design

The 5,000 word task for the WSJ0 Corpus [3] was selected for the ALV evaluation because it represents a well-established LVCSR benchmark within the community and constitutes a good trade-off between computational resources and complexity. The November'92 NIST evaluation set was used for the evaluation data set. Since the original WSJ data was collected at 16 kHz, an 8 kHz downsampled version was created [4]. Processed versions of the data were created to simulate both filtered and additive noise conditions [4]. G.712 filtering was used to simulate the frequency characteristics at an 8 kHz sample frequency and P.341 filtering was used at 16 kHz. A filtered version of the SI-84 training set for the Sennheiser microphone (first channel) was used to construct the first training set, denoted 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, 6 different noise types (car, babble, restaurant, street, airport, and train)

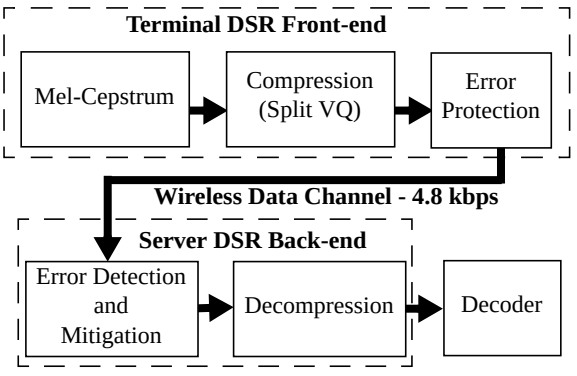

Figure 1: The Aurora standard for a DSR architecture.

were added at randomly selected SNRs between 10 and 20 dB. The goal was an equal distribution of noise types and SNRs. Thus, we had one clean set (893 utterances) and 6 noisy subsets (446 utterances each) for both microphone conditions.

Fourteen evaluation sets (one clean, six noise conditions x two microphone conditions) were defined 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 SNRs between 5 and 15 dB for each of the two microphone types. The SNR averaged across these 12 subsets was designed to be 10 dB.

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 [3]. The pronunciations in the lexicon were extracted from the publicly available CMU dictionary (v0.6) [6] with some local additions. This lexicon is based on a phone set containing 41 phones that includes a short pause and a long pause silence model. A three state left-to-right topology defined the structure of the phonetic models.

The baseline system used in the evaluation was modeled after a 16-mixture WSJ0 system [5] with a WER of 8.3%. Table 1 shows the comparison of this system to the state-of-the-art for a variety of published systems. 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.

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

- Reduced the size of the test set by 50%;
- Adjusted the beam pruning parameters to reduce decoding time by a factor of 6;
- Used only 4 mixtures per state.

The impact of these changes on performance is summarized below in Table 2 and documented extensively in [5].

The baseline recognition system is publicly available [5]. The ETSI WI007 ES 201 108 v1.1.2 front end [1] was chosen for these evaluations. 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.

Site	Acoustic Model Type	Language Model	WER
ISIP	xwrd/gi	bigram	8.3%
CU [7]	xwrd/gi	bigram	6.9%
LT [8]	xwrd/gi	bigram	6.8%
CU [7]	xwrd/gd	bigram	6.6%
UT[9]	xwrd/gd	bigram	6.4%

Table 1: A comparison of performance reported in the literature on the WSJ0 SI-84/Nov'92 evaluation task.

Factor	WER	Relative Degradation
Baseline system (ISIP)	8.3%	N/A
Terminal filtering (ISIP)	8.4%	1%
ETSI frontend	9.6%	14%
Beam adjustments (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 requirements.

3. Analysis

The evaluation of the baseline system on several focus conditions is described below. All experiments were analyzed using the MAPSSWE significance test with $p = 0.1\%$.

3.1 Sample Frequency Reduction

The first focus condition we explored was sample frequency. For Training Set 1 (TS1), degradations due to a reduction in sampling frequency from 16 kHz to 8 kHz did not follow any trend [5]. However, as shown in Figure 2, for Training Set 2 (TS2), statistically significant degradations in performance were observed on the Sennheiser microphone conditions (Test Sets 3-7). Statistically significant test conditions at a 0.1% significance level are indicated by a boldface label.

The overall frequency response of the two microphone conditions is shown in Figure 3. The Sennheiser microphone, as expected, preserves high frequency information better than the second microphone condition, resulting in slightly better performance at a 16 kHz sample frequency. Surprisingly, a similar degradation due to sampling frequency reduction is not observed on perfectly matched conditions (training on TS1 and decoding on Test Set 1) [5], 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 recognition performance. The spectral information provided by low

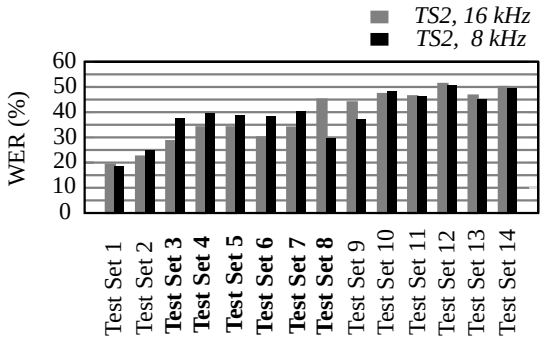

Figure 2: A comparison of the WER for 16 kHz and 8 kHz sample frequencies on TS2.

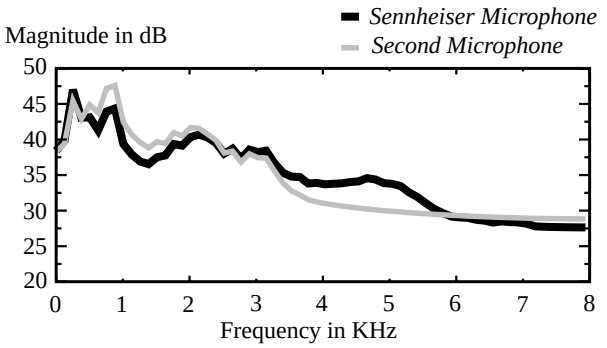

Figure 3: The Sennheiser close-talking microphone preserves frequencies above 3.5 KHz better on the average than the variety of microphones used on the second channel.

frequencies (below 4 kHz) is sufficient to reach the upper bound on performance.

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. 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. The pure silence during training because Training Set 1 consists of only clean data, and hence did not represent a good model of the actual background noise. 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 [5] (a reduction in the number of deletions, rather than insertions, 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 and 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 significantly degrade the performance of the system.

However, there was a significant degradation in performance for five noisy conditions (3, 8, 9, 10, 12) at a 16 kHz sampling frequency and two noisy conditions (7, 11) at an 8 kHz sampling frequency on TS2 [5]. We have not found a consistent explanation as to why these particular noise conditions were adversely affected, but believe it warrants a closer study of the compression algorithm for noisy data.

3.4 Model Mismatch

The best recognition performance was observed on matched training and testing conditions (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 recognition can only be achieved when the test conditions to generate feature vectors are similar 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 for all conditions, 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 Sennheiser microphone and the other half on any one of the 18 microphone types. With the Baum-Welch training algorithm, a

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.

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.

maximum likelihood based parameter estimation method, models trained on TS2 quickly converge towards the Sennheiser microphone in terms of their means and the covariances [10]. Hence, both the clean (Test Set 1) and car (Test Set 8) conditions for the second microphone result in significant degradation in recognition performance, as shown in the second row of the Table 4. Note also that the last three cells in the second row, which correspond to various noise conditions, show less of a degradation in performance than the corresponding conditions in the first row. So there is some value in exposing the models to noise during the training.

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 Figures 4 and 5, we demonstrate that the severity of the degradation in the noisy conditions is reduced by training the models on TS2, which contains samples of the noise conditions. Statistically significant test conditions at a 0.1% significance level are indicated by a boldface label. 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 [5].

On TS1 and TS2, it is observed that performance on the 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 because 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 Aurora large vocabulary evaluation. This system, which is available in the public domain [5], was developed using the standard 5K Wall Street Journal (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 Pentium processor.

We also presented an analysis of the results from these baseline experiments. It is shown that increasing the *sampling frequency* from 8 kHz to 16 kHz results in the 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. In a companion paper, we will present a detailed analysis of the Aurora LV evaluation.

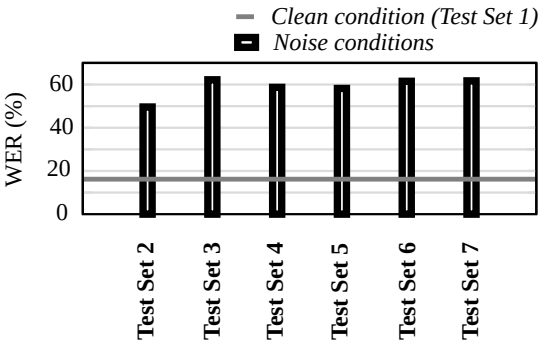

Figure 4: A comparison of the WER for six noise conditions at 8 kHz on TS1.

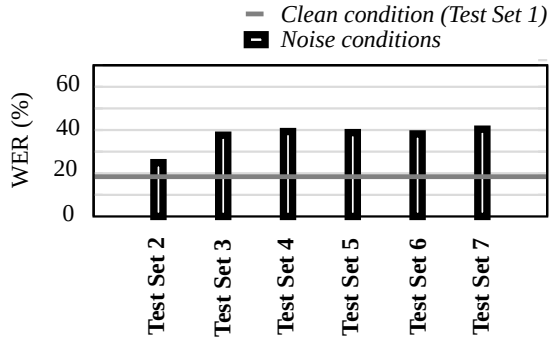

Figure 5: A comparison of the WER for six noise conditions at 8 kHz on TS2.

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] D. Paul and J. Baker, "The Design of Wall Street Journal-based CSR Corpus," *Proceedings of ICSLP*, pp. 899-902, Banff, Alberta, Canada, October 1992.

[4] 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.

[5] N. Parihar and J. Picone, "DSR Front End LVCSR Evaluation," AU/384/02, Aurora Working Group, Dec. 2002 (*http://www.isip.msstate.edu/projects/aurora*).

[6] "The CMU Pronouncing Dictionary," Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 2001 (*http://www.speech.cs.cmu.edu/cgi-bin/cmudict*).

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

[8] 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.

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

[10] R. Sundaram, "Effects of Transcription Errors on Supervised Learning in Speech Recognition," M.S. Dissertation, Mississippi State University, October 2000.