

# Analysis of the Aurora Large Vocabulary Evaluations

*N. Parihar, and J. Picone*

**Institute for Signal and Information Processing**  
**Mississippi State University, Mississippi State, MS 39762 USA**  
{parihar,picone}@isip.msstate.edu

## Abstract

In this paper, we analyze the results of the recent Aurora large vocabulary evaluations. Two consortia submitted proposals on speech recognition front ends for this evaluation: (1) Qualcomm, ICSI, and OGI (QIO), and (2) Motorola, France Telecom, and Alcatel (MFA). These front ends used a variety of noise reduction techniques including discriminative transforms, feature normalization, voice activity detection, and blind equalization. Participants used a common speech recognition engine to postprocess their features. In this paper, we show that the results of this evaluation were not significantly impacted by suboptimal recognition system parameter settings. Without any front end specific tuning, the MFA front end outperforms the QIO front end by 9.6% relative. With tuning, the relative performance gap increases to 15.8%. Both the mismatched microphone and additive noise evaluation conditions resulted in a significant degradation in performance for both front ends.

## 1. Introduction

The Aurora large vocabulary (ALV) evaluations were conducted to standardize an advanced front end (AFE-WI008) for distributed speech recognition applications in a client-server architecture [1]. The ALV evaluations were the second in a series of evaluations designed to promote the development of noise robust speech recognition. The goal for the ALV evaluation was to achieve a 25% relative improvement in word error rate (WER) across a variety of noise conditions compared to the MFCC WI007 front end [1]. The details of the experimental setup and an extended analysis of the baseline system can be found in [1].

A summary of the results presented at the ALV Workshop held in Stuttgart, Germany in February 2002 are shown in Table 1. The overall performance measure for a system was computed as an average of several WERs. First, an average WER was computed across the 14 test sets used in the evaluation for each training condition. Next, the WER for each training condition was averaged. Since the evaluation was conducted at two sample frequencies (8 and 16 kHz), the final WER was the average across both sample frequencies. This number is denoted **Overall WER** in Table 1.

Two consortia participated in the ALV evaluations. The first front end [2] was a collaboration between the CDMA Technologies Group at Qualcomm, the Speech Group at International Computer Science Institute (ICSI), and the Antropic Signal Processing Group at Oregon Health and Science University (OGI). This front end is referred to as the QIO front end and featured three key components: a

15-dimensional MFCC based feature vector generated using data-driven LDA-derived filters, on-line mean and variance normalization, and a multilayer perceptron-based voice activity detector (VAD) [2]. This front end achieved a combined score of 37.5% in the ALV evaluation.

The second contribution [3] resulted from a collaboration between the Human Interface Lab at Motorola Labs, France Telecom R&D, and Alcatel SEL AG (Germany). We refer to this contribution as the MFA front end. It is based on a 12-dimensional MFCC feature vector plus a weighted average of log-energy and the zeroth cepstral coefficient. It employs a two-stage mel-warped Wiener filter for suppressing additive noise [4]. It also incorporates a VAD algorithm that is based on the acceleration of several energy-based measures. Further, it employs a least mean square error blind equalization algorithm for channel normalization. The MFA front end achieved a combined score of 34.5% in the ALV evaluation.

Because these evaluations were conducted using a generic speech recognition system which was not specifically tuned for either front end, it can be argued that the performance achieved by these front ends was suboptimal. Parameters such as the language model scale factor and the word insertion penalty often must be adjusted for a specific front end. Hence, an open issue was whether the ranking of these systems would change if the recognizer was tuned independently for each front end.

Therefore, the main goal of this paper is to explore the sensitivity of the results of this evaluation to parameter tuning.

<b>Baseline MFCC: Overall WER — 50.3%</b>			
8 kHz — 49.6%		16 kHz — 51.0%	
TS1	TS2	TS1	TS2
58.1%	41.0%	62.2%	39.8%
<b>QIO: Overall WER — 37.5%</b>			
8 kHz — 38.4%		16 kHz — 36.5%	
TS1	TS2	TS1	TS2
43.2%	33.6%	40.7%	32.4%
<b>MFA: Overall WER — 34.5%</b>			
8 kHz — 34.5%		16 kHz — 34.4%	
TS1	TS2	TS1	TS2
37.5%	31.4%	37.2%	31.5%

Table 1: The results of the ALV evaluation using a generic baseline speech recognition system (presented at the Feb. 2002 Aurora post-evaluation meeting).

Key recognition parameters, described at length in [5], were optimized independently for each front end. The overall results were then tabulated under these optimized conditions and analyzed across a broad range of noise and microphone-mismatch conditions to assess promising contributions from these front ends.

## 2. Experimental Design

The ALV evaluation is based on the 5,000 word closed-loop WSJ0 task [6], and referred to as the Aurora-4 database [7]. The evaluation was conducted at two sampling frequencies — the original WSJ0 16 kHz data and a downsampled version at 8 kHz. 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. Two training sets (denoted TS1 and TS2), 14 test sets (2 microphone conditions x 7 noise conditions), and 14 short development test sets (2 microphone conditions x 7 noise conditions) were defined for each sampling frequency. The training sets consisted of 7,138 utterances, a short development set consisted of 330 utterances, and an evaluation set consisted of 166 utterances. The construction of these short sets is described in detail in [5].

The two microphone conditions represent the standard conditions supplied with the WSJ Corpus. Parallel recordings were made in which the first channel was always collected with a Sennheiser microphone, while the second channel used one microphone selected from a group of alternative microphones. Seven types of noise were digitally added to the data at specified SNRs. The noise types included street traffic, train stations, cars, babble, restaurants and airports.

The baseline system for the ALV evaluation was a public domain large vocabulary system developed by the Institute for Signal and Information Processing at Mississippi State University in collaboration with the Aurora Working Group [5,8]. The acoustic models consisted of state-tied 4-mixture cross-word triphones. The WSJ0 standard bigram language model (supplied for the 5k closed-loop task) was used to guide a dynamic programming based Viterbi search that uses lexical trees for cross-word decoding. The baseline system, which achieved a WER of 14.0% on the standard 5K WSJ0 task, required 4 xRT for training and 15 xRT for decoding on an 800 MHz Pentium processor. The design and analysis of this system is described in a companion paper [1].

The experimental paradigm to evaluate the influence of front end specific tuning consisted of two sets of experiments. First, all test conditions defined in the ALV evaluation at 8 kHz were repeated for each of the two front ends. For the MFA front end, a binary program was provided by the MFA consortium that included a bug fix for the version submitted to the original evaluation. This bug fix did not affect the overall performance. The QIO consortium provided feature files for all evaluation conditions.

All the system parameters for the first set of experiments were set to the conditions used in the ALV evaluation. The second set of experiments involved individually tuning the system parameters until optimal performance was obtained. The tuning process was executed on the short development test set using Training Set 1 at 8 kHz. These conditions represent matched conditions — clean data recorded using a Sennheiser microphone.

## 3. System Descriptions

The QIO front end [2] is an MFCC-based front end. Fifteen coefficients are used as the base feature set. LDA-derived RASTA filters bandpass filter the temporal trajectories of the log mel-frequency filter bank energies to compensate for the slowly varying convolutional noise introduced due to the channel and the microphone mismatch. On-line cepstral mean subtraction and variance normalization are employed to handle the residual convolutional noise. Each of these channel normalization techniques is known to improve performance independently. A multilayer perceptron (MLP) based VAD eliminates non-speech segments, thereby reducing insertion errors. The input to the MLP consists of three frames of features. The MLP is trained on multiple databases, representing both clean as well as noisy conditions.

Similar to the QIO front end, the MFA front end [4] is also an MFCC-based front end. A 13-dimensional feature vector consists of 12 cepstral coefficients plus log energy. MFA incorporates a two-step noise reduction scheme based on mel-warped Wiener filters to suppress the additive noise. This front end also incorporates an additional waveform SNR weighting block to enhance the SNR of the de-noised signal using a Teager energy operator. A least mean square error based blind equalization is aimed at reducing the mismatch due to channel and microphone variations.

The MFA front end also uses VAD logic. Their approach is based on three measures. The first measure uses the long term acceleration of the energy (computed across the entire spectrum). The second measure uses the acceleration of energy measured over a group of sub-bands of the spectrum likely to contain the fundamental pitch (second, third and fourth filterbank outputs on the mel scale). The third measure uses the variance of the linear-frequency Wiener filter coefficients over the entire frequency band.

The ETSI standard split-vector quantization compression algorithm and framing algorithm are implemented in both the front ends [4]. The bit-stream is decoded, error-detected, error-corrected, and decompressed to form the final features at the back-end server. The delta and acceleration coefficients are computed from the base features at the back end, rather than being transmitted over the channel.

## 4. Results and Analysis

All experiments described below were analyzed using the MAPSSWE significance test [9] with a significance value of 0.1%. The goal for the ALV evaluation was to achieve a 25% relative improvement in word error rate (WER) compared to the MFCC WI007 front end.

### 4.1 Front End-Specific Parameter Tuning

There are four classes of parameters that are most relevant to the tuning performed for this evaluation. Two of these relate to language model and acoustic scores. The language model scale factor controls the relative weight of the language model probabilities compared to the acoustic model probabilities. The word insertion penalty is applied to every word hypothesis and is used to balance insertion and deletion errors. The language model scale factor typically ranges from 5 (for Resource Management) and 20 (for WSJ). The word insertion penalty

QIO	Num states	State Tying Thresh.			LM Sc	Wd Pen	WER (%)
		Spl	Mer	Occ			
Base	3209	165	165	840	18	10	16.1
Tuned	3512	125	125	750	20	10	14.9

Table 2: A comparison of the optimized system parameters to the baseline system parameters for the QIO front end. Beam pruning parameters were set to 300 (state), 250 (model), and 250 (word).

MFA	Num states	State Tying Thresh.			LM Sc	Wd Pen	WER (%)
		Spl	Mer	Occ			
B-line	3208	165	165	840	18	10	13.8
Tuned	4254	100	100	600	18	05	12.5

Table 3: A comparison of the optimized system parameters to the baseline system parameters for the MFA front end.

usually ranges from -10 (Res. Man.) to +10 (WSJ).

The second class of parameters, which have perhaps the most significant impact on performance, relate to the state tying process. The number of tied states can normally be adjusted to improve performance. This parameter balances sparsity and generalization of the data in the decision tree state tying process. We typically reduce the number of states by an order of magnitude. We can also control the degree to which states are merged or split by adjusting parameters related to the likelihood of the state.

In Tables 2 and 3, we show the difference in performance between the baseline system and the tuned system for the QIO and MFA front ends respectively. The tuning process is described in more detail in [5]. Parameter tuning was performed on the matched training condition at 8 kHz (TS1) using the 330-utterance short development test set. The beam pruning parameters (state, model and word) were opened during the tuning process to reduce the influence of pruning.

As shown in Tables 2 and 3, parameter tuning resulted in a small overall improvement — about 1% absolute and 8% relative. The amount of improvement was about the same for both systems — 7.5% relative for QIO and 9.4% relative for MFA. Hence, the ranking of the systems remained the same.

#### 4.2 Detailed Analysis

A more detailed analysis of the results for tuning are shown in Table 4. The pruning beams were scaled back to the values used in the ALV baseline system: 200 (state), 150 (model), and 150 (word). It is observed that the overall relative ranking of the two competitive front ends is not influenced by the tuning process. The average performance of the MFA front end without tuning is better than QIO by 9.6% relative. Front end-specific tuning resulted in an increase in the relative performance gap between the two front ends from 9.6% to 15.8%. While the average performance of the MFA front end remained relatively constant (34.7% to 34.1%), the average performance of the QIO front end dropped by 5.5% relative (38.4% to 40.5%). One possible reason for this drop can be attributed to overfitting of the system parameters on the specific database (matched conditions: TS1 and short devtest set 1) employed for the tuning process.

Significance tests on the 14 test conditions for Training Set 1 (without tuning) showed that the performance of the MFA front end was significantly better than the QIO front end performance on all 14 test conditions. However, on Training Set 2, the MFA front end was significantly better for only Test Sets 5 and 14. Training Set 2 is representative of all noise conditions and includes microphone mismatches. Hence, the TS1 is a good measure for front-end robustness, and perhaps more telling than the matched conditions (TS2). For the ALV evaluations, WERs on the two training sets were weighted equally, thereby decreasing the gap between the two front ends.

In Table 5, we focus on performance for a microphone mismatch training condition. Training Set 1 consisted of clean data recorded with a Sennheiser microphone. Test Set 1 also represents clean data recorded through the same microphone. Test Set 8 represents a mismatched condition since it consists

FE	TS	Tun.	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12	Set 13	Set 14	Avr.
QIO	1	no	17.1	27.2	44.1	47.0	43.1	48.9	44.6	27.5	39.5	49.8	54.9	55.9	52.1	52.0	43.1
QIO	2	no	20.9	22.1	32.8	37.4	35.4	33.6	35.2	24.2	27.4	37.5	42.7	42.2	37.3	41.1	33.6
<b>Average QIO Performance without tuning</b>																	<b>38.4</b>
QIO	1	yes	19.1	31.7	46.8	49.2	45.7	51.1	46.6	30.0	42.2	52.9	55.5	58.3	54.8	55.8	45.7
QIO	2	yes	22.5	23.8	33.6	38.1	36.4	36.2	37.7	25.0	29.5	39.1	44.5	45.0	40.5	41.8	35.3
<b>Average QIO Performance with tuning</b>																	<b>40.5</b>
MFA	1	no	14.5	22.1	37.0	43.2	36.6	43.3	38.2	24.3	29.8	43.4	50.6	48.7	48.6	44.9	37.5
MFA	2	no	18.1	20.6	30.9	36.8	31.6	33.8	31.7	24.3	24.8	34.7	43.3	40.3	38.1	35.7	31.8
<b>Average MFA Performance without tuning</b>																	<b>34.7</b>
MFA	1	yes	14.4	21.5	36.8	42.1	36.5	44.1	36.4	23.3	30.2	43.0	50.2	48.9	47.0	43.6	37.0
MFA	2	yes	16.8	20.7	29.7	36.0	31.0	33.3	32.0	22.5	24.6	34.1	42.3	39.4	37.1	36.1	31.1
<b>Average MFA Performance with tuning</b>																	<b>34.1</b>

Table 4: A detailed analysis of the performance comparison after system-specific tuning.

Train Set	WI007 Baseline		QIO		MFA	
	Set 1 (Sen. Mic.)	Set 8 (Sec. Mic.)	Set 1 (Sen. Mic.)	Set 8 (Sec. Mic.)	Set 1 (Sen. Mic.)	Set 8 (Sec. Mic.)
1	15.4%	36.6%	17.1%	27.5%	14.5%	24.3%

Table 5: A performance comparison for a mismatched microphone condition. Although all the three front ends suffer from significant degradation, the severity is reduced for both the advanced front ends compared to the baseline front end.

of clean data recorded through the second microphone. Though both front ends degraded significantly due to microphone mismatch, this degradation is less severe than the MFCC-based baseline system. The baseline system did not employ any channel normalization techniques such as cepstral mean subtraction.

As shown to the right in Figures 1 and 2, the presence of additive noise resulted in a significant degradation in performance for both the QIO and MFA front ends. The bold labels in these figures represent differences which are statistically significant. This trend is similar to the trend observed on the Aurora MFCC front end based baseline system [1] though the degradations are less severe. The degradation is also less severe when the systems are exposed to noise during training. Performance on the same noisy test sets is much better when training on TS2 because TS2 contains examples of all noise types.

## 5. Summary

In this paper, we have presented a detailed analysis of the results of the Aurora Large Vocabulary evaluation. We have shown that front end specific parameter tuning did not appreciably change the results of the evaluation. The MFA front end still outperformed the QIO front end. In fact, the gap in performance increased slightly after each system was optimized.

It was also shown that mismatched microphones and additive noise significantly degrade recognition performance. Both the QIO and MFA front ends did not degrade as dramatically as the baseline system. In fact, both front ends met the goals set forth in the evaluation — a 25% improvement in performance over the baseline system. Nevertheless, it is clear that there is ample room for new research into ways to make such front ends more robust to unknown noise and microphone mismatch conditions.

## 6. References

- [1] N. Parihar, J. Picone, D. Pearce, and H. G. Hirsch, "Performance Analysis of the Aurora Large Vocabulary Baseline System," submitted to *Eurospeech'03*, Geneva, Switzerland, September 2003.
- [2] C. Benitez, *et al.*, "Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora Task", *Eurospeech'01*, Aalborg, Denmark, September 2001.

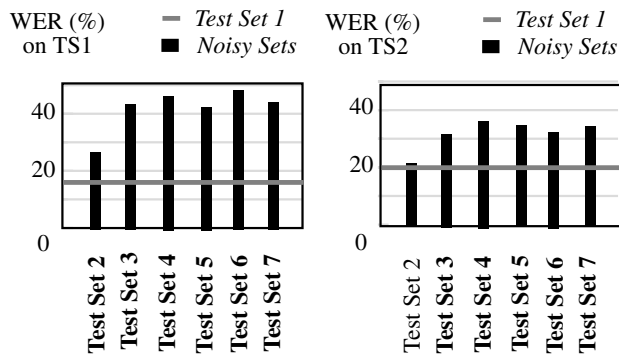


Figure 1: A comparison of the WER for the QIO front end for six noisy conditions.

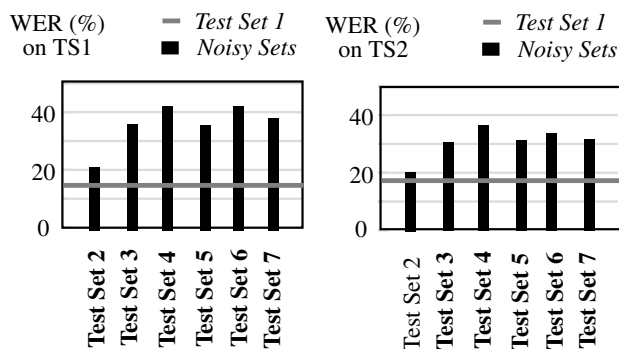


Figure 2: A comparison of the WER for the MFA front end.

- [3] Dusan Macho, *et al.*, "Evaluation of a Noise-robust DSR Front-end on Aurora Databases", *Proceedings of ICSLP*, pp. 17-20, Denver, Colorado, USA, September 2002.
- [4] "ETSI ES 202 050 v1.1.1 Speech Processing, Transmission and Quality aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms," *ETSI*, April 2002.
- [5] N. Parihar and J. Picone, "DSR Front End LVCSR Evaluation," AU/384/02, Aurora Working Group, December 2002 (<http://www.isip.msstate.edu/projects/aurora>).
- [6] 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.
- [7] G. Hirsch, "Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task," *ETSI STQ Aurora DSR Working Group*, December 2002.
- [8] N. Deshmukh, A. Ganapathiraju, J. Hamaker, J. Picone and M. Ordowski, "A Public Domain Speech-to-Text System," *Proceedings of the 6th European Conference on Speech Communication and Technology*, vol. 5, pp. 2127-2130, Budapest, Hungary, September 1999.
- [9] "Benchmark Tests, Matched Pairs Sentence-Segment Word Error (MAPSSWE)," <http://www.nist.gov/speech/tests/sigttests/mapsswe.htm>, Speech Group, National Institute for Standards and Technology, Gaithersburg, Maryland, USA, January 2001.