PERFORMANCE ANALYSIS OF ETSI FRONTEND IN DSR FRAMEWORK[1]

N. Parihar, and J. Picone

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

D. Pearce

European Telecommunications Standards Institute
650, Route des Luciole
06921 Sophia Antipolis CEDEX
France
bdp003@motorola.com

H. G. Hirsch

Niederrhein University of Applied Sciences
Reinarzstrasse
Krefeld
Germany
hirsch@hs-niederrhein.de

ABSTRACT

In this paper, we present the analysis of the baseline experiment results for Aurora Evaluations, designed to calibrate the influence of the six focus conditions on the speech recognition performance for the ETSI frontend, in a Distributed Speech Recognition architecture. The results show that increasing the *sampling frequency* from 8 kHz to 16 kHz resulted in the significant improvement in the performance only for noisy conditions. Similarly, *utterance detection* resulted in significant improvement only on noisy conditions. Moreover, these improvements due to endpointing are not visible with multi-condition training. The lossy VQ based *compression* algorithm did not result in any significant degradation. *Model Mismatch* due to mismatched training and testing conditions resulted in significant degradations in the performance. Similarly, the best performance is seen on matched *microphone conditions*. The *six test noisy conditions* degraded performance, even when the training conditions were exposed to similar noise conditions.

1. INTRODUCTION

With the increase in compute power of the miniature mobile devices such as cell phones, standardization of the communication protocols, advancement of the ASR technology, and popularity of the mobile devices, speech recognition has become a standard application provided on the mobile devices such as cell-phones. One of the common architectures popular for such applications is the Client/Server Distributed Speech Recognition architecture. The Aurora Standard for DSR [1] is shown in Figure 1. The features are extracted, compressed, framed using an standard error-detection-and-correction algorithm on the client and then transmitted over the noisy channel. These features are then recovered at the server side, and used for recognition. The key advantage of this approach being the ability to extract features on terminal devices with tiny compute power in real time. Because these features are used for recognition on the server-end, sophisticated noise-robust algorithms and recognition algorithms can be employed to boost performance.

In this paper, we discuss and analyze the performance of the ETSI frontend [1] on six focus conditions within the DSR framework as defined by Aurora Working Group.

Sample Frequency Reduction: Two sampling rates, 16 kHz and 8 kHz, were evaluated.

Utterance Detection: The training and evaluation utterances were exercised to study the effect of endpointing on the overall WER on the various noisy conditions.

Compression: A vector quantization-based compression scheme which compresses features to 4800 b/s was used to evaluate the degradation in performance.

Model Mismatch: Mismatched training and testing conditions are inevitable in a rapidly evolving area such as wireless communications. A simple evaluation was conducted to calibrate this condition.

Microphone Variation: Both the microphone conditions available in the WSJ0 task [2] were evaluated.

Additive Noise: Six noise types collected from street traffic, train stations, cars, babble, restaurants and airports at varying signal-to-noise ratios were calibrated.

2. EXPERIMENTAL SET-UP

All the experiments employed state-tied cross-word speaker-independent triphone acoustic models with four Gaussian mixtures per state. A single pass ngram decoding based on dynamic-programming search guided by a standard bigram language model for WSJ0 task [2] was performed. All the

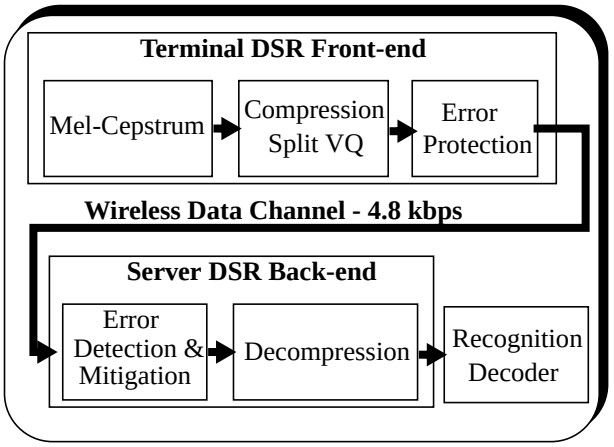

Figure 1: Aurora Standard for DSR architecture.

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 the ETSI.

defaults parameters were set to the default setting of our best 16 mixture WSJ0 system [4]. All the MAPSSWE significance tests [7] were conducted at a significance level of 0.1%.

Front-end: The ETSI frontend *ES 201 108 v1.1.2* was chosen as the baseline for Aurora Evaluations [4]. It is a standard MFCC front-end that extracts the 13 cepstral coefficients and log-energy to form a 14 dimensional feature vector for each frame. The frame length and window size is 10 msecs and 25 msecs, respectively. These feature vectors are compressed though a standard lossy vector quantization algorithm to reduce the transmission rate to 3800 b/s. At the server-end, the 14-dimensional feature vector is reconstructed using the standard decompression algorithm. For all the experiments, the zeroth order cepstral coefficient was thrown to form the 13 dimensional feature vectors. The delta and double-delta coefficients were computed on the fly to get 39 dimensional vectors.

LVCSR Toolkit: The state-of-the-art public domain ISIP Prototype system [5] was used for the experimentation. This toolkit efficiently and transparently handles tasks of varied complexity, from connected digits to spontaneous conversations.

Database: The DARPA WSJ0 [2] has two-channel recordings of the same utterances that were made at 16 kHz. Channel 1 consists of the same microphone for all speakers — a Sennheiser HMD-414 close-talking microphone. The second channel included a sampling of 18 different types of microphones. A downsampled version of this task at 8 kHz was created [3].

The November 92 NIST evaluation set is a speaker-independent set consisting of 330 utterances. Due to time constraints and the large number of experiments that needed to be run in order to evaluate various conditions, we decided to reduce the 330 utterance eval set to a 166 utterance eval subset [4].

The processed versions of the training and evaluation utterances were generated to simulate both the filtered and additive noise conditions [3]. The G.712 filtering was used to simulate the frequency characteristics at an 8 kHz sample frequency and P.341 filtering was used for simulation at 16 kHz. The Training Set 1 consists of the filtered version of the SI-84 training set (7138 utterances) recorded with the Sennheiser microphone.

For Training Set 2, the filtered 7138 training utterances are divided into two blocks: 3569 utterances (half) recorded with the Sennheiser microphone, and the remaining half recorded with a different microphone (18 different microphone types were used). No noise is added to one-fourth (893 utterances) of each of these subsets. To the remaining three-fourths (2676 utterances) of each of these subsets, 6 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 of noise types and SNRs. Thus, we had one clean set (893 utterances) and 6 noisy subsets (446 utterances each) for both the microphone conditions.

Fourteen evaluation sets were defined in order to study the degradations in speech recognition performance due to microphone conditions, and noisy environments. Each of the filtered versions of the evaluation set recorded with Sennheiser microphone and second microphone were selected to form the two eval sets. The remaining 12 subsets were defined by randomly adding each of the 6 noise types at randomly chosen SNR between 5 and 15 dB for each of the two microphone types. The goal was to have an equal distribution of each of the 6 noise types and the SNR with an average SNR of 10 dB.

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 the phone set containing 41 phones that includes the sp and sil. A three emission state left-to-right topology defined the structure of the phonetic models.

3. ANALYSIS

3.1 Sample Frequency Reduction

For Training Set 1, degradations due to a reduction in sampling frequency from 16 kHz to 8 kHz did not follow any trend [4]. However, for Training Set 2, statistically significant degradations in performance were observed on the Sennheiser microphone conditions (Test Sets 3-7) [4]. The performance at four sample conditions on both the Training sets is shown in Table 1 and 2. The overall frequency response of the two microphone conditions on the speech data of a typical utterance as shown in Figure 2 demonstrates that the Sennheiser microphone preserves high frequency information better than the second microphone condition.

However, no significant degradation due to sampling frequency reduction is observed on matched conditions (Sennheiser Microphone) — training on Training Set 1 and decoding on Test Set 1. The additional information provided by high frequencies (between 4 kHz and 8 kHz) does not contribute to any additional improvement in recognition performance.

3.2 Utterance Detection

The utterance detection resulted in a significant improvement in performance on Test Sets 2-14 when the system was trained on Training Set 1 [4]. Two sample test conditions in Table 3 show that the reduction in insertion errors is primarily responsible for improvement in the performance. In this case, the “silence” model learned only 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 the non-silence words, resulting in insertion type errors.

In contrast, for Training Set 2, a significant improvement in performance was detected only for Test Set 8 [4]. A reduction in the number of deletions, rather than insertions, was primarily responsible for this improvement. In other words, 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 Training Set 1 for both the

sampling frequencies [4]. Because there is no significant degradation for Test Set 1 which is a matched condition, it is natural to 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 Training Set 2 [4]. We have not found a consistent explanation as to why these particular noise conditions were adversely affected.

3.4 Model Mismatch

The best recognition performance was observed on matched training and testing conditions (Training Set 1 and Test Set 1), when all the utterances were recorded with a Sennheiser microphone, as shown in Figure 3. For all other conditions involving Training Set 1, the recognition performance degraded significantly. Systems trained on Training Set 2 performed significantly better than those trained on Training Set 1 across all noise conditions. These trends were consistent for both the sampling frequencies and both the compression conditions [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. If there are consistent differences in SNR, background noise, or microphone, there will be a significant degradation in performance without any adaptation scheme.

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 Training Set 1, 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 the word error rate, 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).

Training Set 2 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 maximum likelihood based parameter estimation method, this fact implies that models trained on Training Set 2 quickly converge towards the Sennheiser microphone in terms of their means and the covariances [8]. 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 the noise conditions and at both the sample frequencies. However, the severity of this degradation can be limited by exposing the models to noise conditions during the training process. In Figures 3 and 4, we demonstrate that the severity of the degradation in the noisy conditions is reduced by training the models on Training Set 2, which contains samples of the noise conditions. 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].

4. SUMMARY

In this paper, we presented the analysis of the results of the baseline experiments for Aurora evaluations aimed at calibrating the performance of the ETSI frontend on the six focus conditions — sample frequency reduction, utterance detection, compression, model-mismatch, microphone variation, and additive noise. This analysis provides the additional useful insights in recognition performance in a DSR framework. For example, the higher 16 kHz sampling frequency results in significant performance improvement only on noisy conditions when clean data is used for training. Similarly, the utterance detection results in significant improvement in the performance due to elimination of the insertion type errors. The future work will include tuning the decoder parameters using the ETSI frontend with an aim to calibrate the influence of tuning the ETSI frontend on recognition performance.

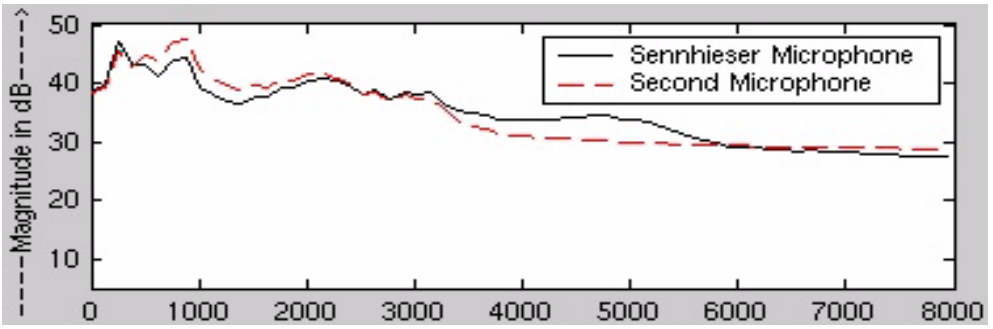

Figure 2: Comparison of the magnitude of the frequency response of the Sennheiser microphone and the second microphone derived from the speech segment from the utterance id *441c020b*, digitized at 16 KHz. The high quality Sennheiser microphone preserves the frequencies above 3.5 KHz while the second microphone filters the high frequencies.

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. Paul and J. Baker, "The Design of Wall Street Journal-based CSR Corpus," *Proceedings of the International Conference on Spoken Language Systems (ICSLP)*, pp. 899-902, Banff, Alberta, Canada, October 1992.

[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," *http://www.isip.msstate.edu/publications/reports/aurora_frontend/2002/report_012202_v21.pdf*, Aurora Working Group, December 2002.

[5] N. Deshmukh, A. Ganapathiraju, J. Hamaker, J. Picone and M. Ordowski, "A Public Domain Speech-to-Text System", *6th European Conference on Speech Communication and Technology*, Vol. 5, pp. 2127-2130, Budapest, Hungary, September 1999.

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

[7] "Benchmark Tests, Matched Pairs Sentence-Segment Word Error (MAPSSWE)," *http://www.nist.gov/speech/tests/sigtests/mapsswe.htm*, Speech Group, NIST, USA, January 2001.

[8] R. Sundaram, "Effects of Transcription Errors on Supervised Learning in Speech Recognition," M.S. Dissertation, Institute for Signal and Information Processing, Mississippi State University, October 2000.

Sample Frequency	Set 1	Set 3	Set 8	Set 10
16 kHz	14.0%	**57.2%**	**52.7%**	74.3%
8 kHz	16.2%	**62.2%**	**37.4%**	69.8%

Table 1: A comparison of the WER for *16 kHz and 8 kHz* sample frequencies for *Training Set 1* for 4 sample test sets. Test set conditions which are statistically significant at a 0.1% significance level are indicated by a boldface label.

Sample.Frequency	Set 1	Set 3	Set 8	Set 10
16 kHz	19.2%	**28.5%**	**45.0%**	47.2%
8 kHz	19.2%	**37.6%**	**29.7%**	48.3%

Table 2: A comparison of the WER for *16 kHz and 8 kHz* sample frequencies for *Training Set 2* for 4 sample sets. Test set conditions which are statistically significant at a 0.1% significance level are indicated by a boldface label.

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: A comparison of experimental results for *without and with endpointed* data for *Training Set 1* at *16 KHz* with *no feature vector compression*. The two sample test sets demonstrate that reduction in insertion type errors are primarily responsible for the significant reduction in WER.

Training Set	Set 1	Set 8	Set 2	Set 9
1	16.2%	37.4%	49.6%	59.7%
2	18.4%	29.7%	24.9%	37.3%

Table 4: A significant performance degradation occurs for the *second microphone condition* on both training sets. On Training Set 1, the models are representative of the Sennheiser microphone condition and hence, the performance drops significantly due to model mismatch on the remaining conditions. The models trained on Training Set 2, which have still been exposed to significant amounts of the Sennheiser data, do slightly better on the other noise conditions.

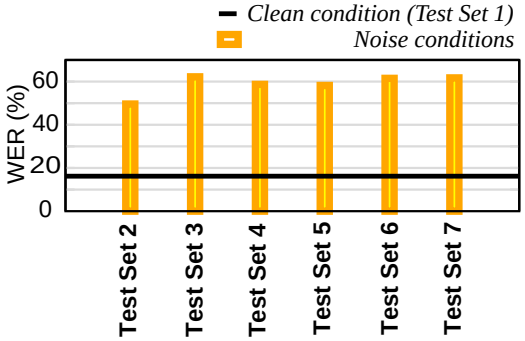

Figure 3: Comparison of the WER for six noise conditions at *8 kHz*. *Training Set 1* was used for Training in this case. Statistically significant test conditions are indicated by a boldface label.

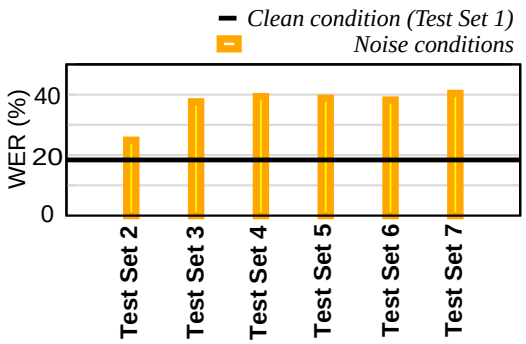

Figure 4: Comparison of the WER for six noise conditions at *8 kHz*. *Training Set 2* was used for Training in this case. Statistically significant test conditions are indicated by a boldface label.