GENERAL ISSUES IN ENVIRONMENTAL NOISE TRACKING FOR ROBUST IN-VEHICLE SPEECH APPLICATIONS: SUPERVISED vs UNSUPERVISED ACOUSTIC NOISE ANALYSIS

Murat Akbacak and John H.L. Hansen

Robust Speech Processing Group
Center for Spoken Language Research
University of Colorado at Boulder
Boulder, CO, 80302, U.S.A
{murat,jhlh}@cslr.colorado.edu   Web: http://cslr.colorado.edu

ABSTRACT

In this paper, we present an overview of *Environmental Sniffing* [1, 2] framework with current extensions to the system. The framework of Environmental Sniffing is focused on detection, classification and tracking changing acoustic environments. Here, we extend the framework to detect and track acoustic environmental conditions which are determined in an unsupervised approach as opposed to the supervised approach employed in [1, 2]. Knowledge extracted about the acoustic environmental conditions is used to determine which environment dependent speech recognizer to use. Critical Performance Rate (CPR), previously considered in [1, 2], is also presented. The sniffing framework is compared to a ROVER solution for automatic speech recognition (ASR) using different noise conditioned recognizers in terms of Word Error Rate (WER) and CPU usage. Results are presented in this paper for supervised noise analysis. Results show that the model matching scheme using the knowledge extracted from the audio stream by Environmental Sniffing does a better job than a ROVER solution both in accuracy and computation. A relative 11.1% WER improvement is achieved with a relative 75% reduction in CPU resources.

1. INTRODUCTION

Significant advances in ASR technology have been achieved in applications where the environmental noise condition is constant. Most recently, ASR research focus has shifted to real-world environments where changing environmental noise conditions represent significant challenges in maintaining ASR performance.

All efforts in the field of noisy speech recognition have been directed at reducing the mismatch between training and operating conditions such as speech enhancement, noise resistant features, re-training and multi-style training, and model adaptation. Each solution has both advantages and disadvantages [3].

Today, state of the art ASR systems use a parallel bank of recognizers in a ROVER paradigm [4] to take advantage of the methods mentioned above. This framework seeks to reduce word error rates for ASR by exploiting differences in the nature of the errors made by multiple speech recognizers which use different features in the feature extraction step, different noise compensation schemes in the enhancement step, or different model adaptation schemes. The disadvantage of this framework is the high computational power it requires. There are also many open questions: for example how important is the combination order of the system hypotheses?, which recognizers should be used?, how many systems should we combine?, is it advantageous to preprocess or normalize the systems' outputs prior to combination? Most researchers take the approach of using more recognizers, and try to make each ASR engine different in some meaningful way (i.e., different features, trained in different noise, etc.) to leverage the potential differences in recognition errors.

In [1, 2], we addressed the problem of changing acoustic environmental conditions in speech tasks by proposing a new framework entitled *Environmental Sniffing* to detect, classify and, track changing acoustic environmental conditions and extract knowledge about the environmental noise. The goal is to do smart tracking of environmental conditions and direct the ASR engine to use a solution specific to each environmental condition. In our previous studies[1, 2], a supervised training process with pre-defined noise types in car environment was used. An alternative approach would be to consider an unsupervised training method with no prior noise type list. To do this, here, we use BIC (Bayesian Information Criteria) based segmentation and clustering to obtain noise entries without associated physical events. In applications (e.g., digital archives, etc.) where it is not feasible to manually segment and label the different noise types (i.e., large number of noise types might result in inconsistent human labeling), unsupervised acoustic noise analysis gains importance.

The organization of our paper is as follows. In Section 2, we present a background on noisy speech recognition. In Section 3, we specialize the general framework of sniffing environmental noise for an in-vehicle hands-free digit recognition task. In Section 5, we present the unsupervised noise analysis module that consists of BIC based segmentation and clustering. In Section 4, algorithm formulation of environmental sniffing in a noisy-speech scenario is presented. Section 6 includes the formulation of the critical performance rate (CPR) of Environmental Sniffing for the digit car task. In Section 7, evaluations of the framework integrated into an in-vehicle ASR engine is presented. Section 8 discusses some further research issues for sniffing with conclusions given in Section 9.

This work was supported in part by DARPA under Grant No. N66001-8906 and NSF Cooperative Agreement No. IIS-9817485

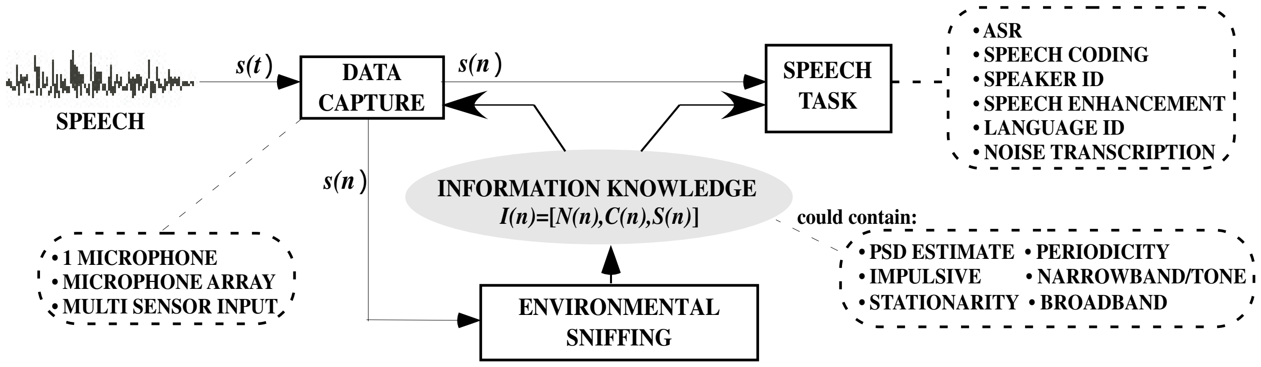

Fig. 1. How a proposed Environmental Sniffer works.

2. BACKGROUND ON ROBUST SPEECH RECOGNITION

All efforts in the field of noisy speech recognition have been directed at reducing the performance mismatch between training and operating conditions. These techniques are grouped into the following categories:

- Re-Training & Multi-Style Training
- Speech Enhancement & Feature Enhancement
- Noise Resistant Features
- Model Adaptation

In the re-training method, an "environment-dependent" system is re-trained with data from new testing environments. The main disadvantage of this technique is the lack of *a priori* knowledge of environmental characteristics. In addition to this, data collection and transcription is time consuming and the training process is extremely computationally expensive.

In multi-style training, an "environment-independent" system is trained by pooling data from different acoustical environments. The disadvantage of this method is the lack of sufficient environments needed to achieve environment independence. Also, it is unclear how speech from diverse environmental conditions contributes to the overall speech recognition model.

Most early work towards robustness has been derived from classical techniques developed in the context of speech enhancement ([5] offers a good historical summary, and [6] represents a more recent summary on enhancement techniques). The goal is to transform noisy speech into a reference environment, and recognize it with a system trained in the reference environment. Speech enhancement methods offer the distinct advantage of requiring no training data, and can be enabled or disabled with limited changes to the subsequent speech task.

In the robust features method, it is assumed that the system is noise independent, and uses the same system configuration for both noisy and clean speech recognition. The goal is to derive noise resistant parameters. One of the advantages of this technique is that in general weak or no assumptions are made about the noise (i.e., no explicit estimation of the noise statistics is required). On the other hand, this could be a shortcoming since it is impossible to make full use of characteristics specific to a particular noise type.

Model adaptation schemes transform the speech models created in the reference environment in order to accommodate the evolving noisy environment. Since accurate estimates of the noise statistics are required, this method can be sensitive to varying SNR (signal-to-noise ratio) and non-stationary noise environments.

While speech or feature enhancement, robust features, or model adaptation can be effective for robust speech recognition in noise, it is difficult to outperform a system that is trained in the same noise type and level when noisy conditions are stationary. While speech enhancement and model adaptation methods typically have access to a short segment of noise for statistical characterization, a full re-training approach typically requires several hours of speech in noise data. As such, many ASR researchers have migrated towards dedicated trained systems.

In more recent studies, as computational power has increased with the help of high-speed computers, a parallel bank of recognizers has been used in a Recognizer Output Voting Error Reduction (ROVER) paradigm [4]. This method seeks to reduce word error rates for dedicated ASR engines by exploiting differences in the nature of the errors made by multiple speech recognizers which use different features in the feature extraction step, different noise compensation schemes in the enhancement step, or different model adaptation schemes. The disadvantage of this method is the high computational power it requires, making it less feasible for real-time or dialog applications (i.e., it is not uncommon for a system to run 100 times slower than real-time). In addition, it is not a general solution for other speech systems (e.g., speech coding, speech enhancement, etc.).

3. SYSTEM ARCHITECTURE

A proposed general system architecture diagram for Environmental Sniffing [1] is shown in Fig. 1. Digitized speech is denoted as $s(n)$, captured from an input sensor (i.e., single or multi-microphone) and acoustic environmental information as $I(n)$ which is a function of the input signal.

In a sample scenario, $s(n)$ may be the audio data obtained in a vehicle with a microphone array, the speech task may include model adaptation within an ASR system, and $I(n)$ may consist of the existing noise types with time tags and the power spectral estimates of the environmental noise with a stationarity measure. Here, $I(n)$ could also contain a suggestion to use one of several adaptation schemes (Jacobian adaptation [7], MLLR [8], PMC [9], etc.), or alternative parameterization (MFCC, LPC, PLP [10], SBC [11], WPP [11], etc.) which gives the best performance for the environmental noise knowledge estimated through Environmental Sniffing.

In addition to environmental noise knowledge $N(n)$, $I(n)$ may contain $S(n)$- knowledge about the speaker identity and $C(n)$-channel information for the speech tasks having multi-*SEC* characteristics. Knowledge from $S(n)$ can be used to monitor the speaker's speaking style. It may consist of the accent and stress levels or emotion of the speaker so that correct pronunciation and duration modeling, or acoustic model compensation techniques can be employed [12]. This knowledge may be used in speech coding to improve the naturalness of speech. $C(n)$ may provide the knowledge of channel type (bandwidth, type of distortion, etc.), impact of channel (fade-out, channel bias, etc.) to improve the ASR system performance by having reliable parameters for feature enhancement or model adaptation.

In the remainder of this study, we focus on changing acoustic environmental conditions and construct the *Environmental Sniffing* framework to extract environmental noise knowledge $N(n)$ from an input audio stream. In other words, $I(n)$ will consist of only environmental noise knowledge $N(n)$, with a constant channel $C(n)$ and speaker $S(n)$ traits.

Fig.2 shows a proposed robust ASR system for in-vehicle route information originally presented in [15].

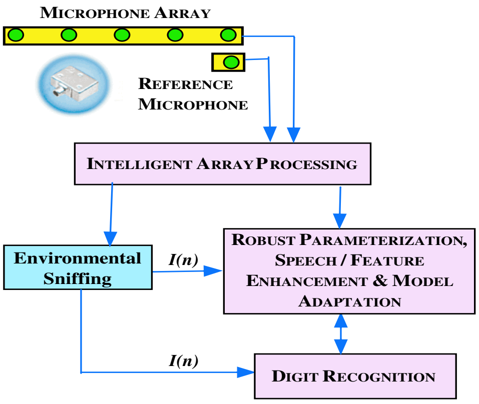
MICROPHONE ARRAY
REFERENCE MICROPHONE
INTELLIGENT ARRAY PROCESSING
Environmental Sniffing
I(n)
ROBUST PARAMETERIZATION, SPEECH / FEATURE ENHANCEMENT & MODEL ADAPTATION
I(n)
DIGIT RECOGNITION

Fig. 2. An in-vehicle digit recognition system.

The motivation for selecting this environment is the huge diversity of acoustic environmental conditions and the need to maintain near real-time performance for route navigation dialogs.

In Fig. 2, we see that environmental sniffing plays a central role in determining the environment information which could be used to direct front-end array processing, parameterization, speech enhancement, model adaptation, or ASR model selection for effective speech recognition. Therefore, the environmental sniffer could be a passive system and simply provide information $I(n)$ to any prior or subsequent speech processing tasks. In contrast, the sniffer could instead take control and direct appropriate microphone array processing, feature selection/processing, and/or adjust model adaptation depending on the environmental knowledge and confidence. For the purpose of this paper, the Environmental Sniffing framework will be employed for ASR model selection.

4. ENVIRONMENTAL SNIFFING

In [1], we focused on extracting knowledge concerning the acoustic environmental noise using a noise-only audio database containing 8 noise conditions in a car environment. In [2], we presented a broad class monophone recognition based system for sniffing noisy-speech data, as shown in Fig. 3. In addition to 8 noise conditions, the acoustic condition set contains also the clean condition-CL as shown in Table 1.

Acoustic Condition Set		
1	N1	Idle noise consisting of the engine running with no movement and windows closed
2	N2	City driving without traffic and windows closed
3	N3	City driving with traffic and windows closed
4	N4	Highway driving with windows closed
5	N5	Highway driving with windows 2 inches open
6	N6	Highway driving with windows half-way down
7	N7	Windows 2 inches open in city traffic
8	NX	Others
9	CL	Clean (i.e., noise-free)

Table 1. In-vehicle acoustic conditions considered.

After defining a set of broad phone classes (e.g., STP- stop, FRC- fricative, NSL- nasal, VWL- vowel, SIL- silence, etc.), an HMM is trained for each *(broad phone class, acoustic condition)* pair. As an example, an HMM for the pair (FRC,N1) is trained from a clean database of fricatives degraded by acoustic condition $N1$. These acoustic models are used during the broad class monophone recognition.

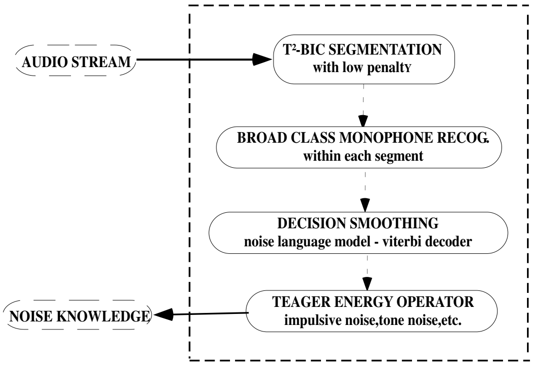
AUDIO STREAM
T²-BIC SEGMENTATION with low penalty
BROAD CLASS MONOPHONE RECOG. within each segment
DECISION SMOOTHING noise language model - viterbi decoder
TEAGER ENERGY OPERATOR impulsive noise,tone noise,etc.
NOISE KNOWLEDGE

Fig. 3. Environmental Noise Sniffing.

As Fig. 3 shows, the incoming audio stream is first segmented into acoustically homogeneous speech blocks using our T^2-BIC [14] segmentation scheme with a low false alarm penalty (i.e. false alarms are tolerable to ensure we capture all potential marks, both true and false). For each segment, a lattice is generated in an FST (Finite State Transducer) format via phoneme recognition. During decision smoothing, the resulting phone-lattice of each segment is combined with an FST representing the noise language model. The costs of noise transitions in the FST representing the noise language model is inversely proportional with the transition probabilities presented in [1].

5. UNSUPERVISED NOISE ANALYSIS

In [1, 2], we chose to employ supervised training using human transcribed (e.g., pre-defined noise types) noise data for our evaluations. We did this in order to tag noise events for in-vehicle speech dialog systems. An alternative approach is to consider an unsupervised training method with no prior noise type list.

In our Environmental Sniffing framework, we propose to use BIC based segmentation and clustering to obtain noise entries without associated physical events. After extracting homogeneous noise segments via BIC based segmentation, we use agglomerative bottom-up BIC based clustering to merge acoustically similar noise segments.

Assuming that each acoustic homogeneous noise block is modeled as one multivariate Gaussian process, we can consider the audio stream as two nested models : M where $X = \{x_i | i = 1, 2, \dots, N\}$ is independent and identically distributed as a single Gaussian $N(\mu, \Sigma)$, and M_2 where the initial frames $\{x_i | i = 1, 2, \dots, b\}$ are drawn from one Gaussian $N(\mu_1, \Sigma_1)$ while the remaining frames $\{x_i | i = b+1, b+2, \dots, N\}$ are drawn from another Gaussian $N(\mu_2, \Sigma_2)$. Using this representation, the BIC difference between the two models is found as,

$$\begin{aligned} \triangle BIC_b &= \frac{1}{2}(N \log \mid \Sigma \mid - b \log \mid \Sigma_1 \mid - (N-b) \log \mid \Sigma_2 \mid) \\ &\quad - \frac{1}{2}\lambda(d + \frac{1}{2}d(d+1)) \log N, \qquad (1) \end{aligned}$$

where λ is the penalty factor to compensate for small sample size cases, and d is the cepstral feature dimension.

Initially, each segment is a cluster by itself and the clusters are modeled by a single Gaussian. At each step of the clustering algorithm, a similarity measure is calculated for each pair of clusters. The two closest clusters are merged if the corresponding BIC variation, given by Eq.1, is negative. If the difference is positive, the algorithm is stopped.

6. CRITICAL PERFORMANCE RATE

In [1], we defined a critical performance rate (CPR) in a general sense. In [2] we specialized the formulation of CPR to a specific case where Environmental Sniffing framework is used for model selection within an ASR system. The Environmental Sniffing framework determines the initial acoustic model to be used according to the environmental knowledge it extracts. The knowledge in this context, will consist of the acoustic condition types with time tags. Following is the formulation of CPR revisited:

Let us denote the error matrix for noise classification as ϵ:

$$\epsilon = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1N} \\ e_{21} & e_{22} & \dots & e_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ e_{N1} & e_{N2} & \dots & e_{NN} \end{bmatrix}. \qquad (2)$$

For $i = j$, $1 \leq i, j, \leq N$, e_{ij} is zero, and for $i \neq j$, e_{ij} is the classification error rate (in a range 0-1) for the error type where the i^{th} noise class is classified as the j^{th} noise class.

Assume that there are N initial acoustic models [1] to be used during recognition, each corresponding to an environmental condition. These models can be trained by simply re-training HMMs for N different acoustic conditions. Assume that there is enough diversity among noise conditions so that for a noise type during decoding, using the matched acoustic model as an initial model during model adaptation yields the lowest WER. Let us define a matrix W as follows:

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1N} \\ w_{21} & w_{22} & \dots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \dots & w_{NN} \end{bmatrix} \qquad (3)$$

where w_{ij} represents the WER value for the case where test tokens are from the i^{th} noise class, but the j^{th} acoustic model which is trained from the j^{th} acoustic condition is used as an initial model. Using the matrix W, we can assign a cost value for each error type so that each error rate e_{ij} can be weighted by the normalized cost values to calculate the Critical Performance Rate (CPR) of the Environmental Sniffing framework. For the error type where the i^{th} noise class is classified as the j^{th} noise class, the cost is $\Delta w_{ij-ii} = w_{ij} - w_{ii}$, which is the performance deviation of the ASR engine by using the j^{th} acoustic model during decoding instead of using the correct i^{th} acoustic model.

Since some noise conditions occur more frequently than others, each noise condition will have an *a priori* probability denoted as follows:

$$\vec{\mathbf{a}} = a_1 \quad a_2 \quad \dots \quad a_N \qquad (4)$$

Now, we can formulate the Critical Performance Rate as:

$$\begin{aligned} CPR &= 1 - \sum_{i=1}^{N} a_i \sum_{j=1}^{N} \frac{\Delta w_{ij-ii}}{n_{ij}} e_{ij} \\ &= 1 - \sum_{i=1}^{N} a_i \sum_{j=1}^{N} \frac{w_{ij} - w_{ii}}{\frac{\sum_{k=1, k \neq i}^{N} w_{ik} - (N-1) w_{ii}}{N-1}} e_{ij} \\ &= 1 - \sum_{i=1}^{N} a_i \sum_{j=1}^{N} \frac{w_{ij} - w_{ii}}{\frac{\sum_{k=1}^{N} w_{ik} - N w_{ii}}{N-1}} e_{ij} \\ &= 1 - \sum_{i=1}^{N} a_i \sum_{j=1}^{N} C_{ij} e_{ij} \qquad (5) \end{aligned}$$

where n_{ij} is the normalization term and C_{ij} is the normalized cost value for the error type where the i^{th} noise class is classified as the j^{th} noise class.

In matrix form, Eq. 5 becomes:

$$CPR = 1 - diag\left\{\mathbf{C} \cdot \epsilon^{\mathbf{T}}\right\} \cdot \vec{\mathbf{a}}^{\mathbf{T}} \qquad (6)$$

where $\mathbf{C}$ is the normalized cost matrix having entries C_{ij}.

If all noise conditions have equal *a priori* probabilities $1/N$, and all error types have equal costs (e.g., each error type has the same impact on the subsequent system's performance) then we obtain

$$CPR = 1 - \sum_{i=1}^{N} \frac{1}{N} \sum_{j=1}^{N} e_{ij} \qquad (7)$$

The goal, in terms of performance, is to optimize the critical performance rate rather than optimizing the environmental noise classification performance rate, since it is more important to detect and classify noise conditions that have a more significant impact on ASR performance.

[1] If there are M ($M \leq N$) initial models, the $\mathbf{W}$ matrix will still be NxN, since some noise classes will use the same acoustic model, and the cost of errors among these noise classes will be zero.

We can use Eq. 5 by setting $n_{ij} = 1$ (without normalizing the costs) to calculate $1 - CPR$, which will actually be the expected performance deviation from the highest achievable performance when we use Environmental Sniffing. In our example, using the matched acoustic model for an environmental condition will achieve the lowest WER. Using the expected performance deviation, we can estimate the performance of model matching employing the Environmental Sniffer.

7. EVALUATIONS

In our evaluations, we degraded the TI-DIGIT database at random SNR values ranging from -5 dB to +5 dB (i.e., -5,-3,-1,+1,+3,+5 dB SNR) with 8 different in-vehicle noise conditions using the noise database from [1]. A 2.5-hour noise data set was used to degrade the training set of 4000 utterances, and the 0.5 hour set was used to degrade the test set of 500 utterances (i.e., open noise degrading condition). Each digit utterance was degraded with only one acoustic noise condition. Next, an HMM was trained for each *(broad phone class, acoustic condition)* pair. The phoneme classes for digits were mapped to 7 broad classes (including SIL- silence). Since we have 9 acoustic conditions (including CL- clean) in the acoustic condition set, at the end we had 7 x 9 = 63 HMMs. Each *(broad phone class, acoustic condition)* was listed in the lexicon during decoding. A total of 7 silence models were also used as filler models.

Each digit utterance was decoded into a sequence of *(broad phone class, acoustic condition)* pairs. At each leaf of the lattice, the *(broad phone class, acoustic condition)* pair was mapped to the corresponding acoustic condition (e.g., STP-N1 was mapped to N1). The resulting lattice in FST format is combined with the lattice representing the noise language model to find the most likely noise sequence.

For acoustic model training and decoding, we used CSLR's Large Vocabulary Continuous Speech Recognizer SONIC [16]. AT&T's FSM Toolkit [17] was used to combine the phone-lattice and the noise language model.

7.1. Sniffing Results:

Using the sniffing framework presented in Sec. 4, each utterance was assigned to an acoustic condition. Using the fact that there was only one acoustic condition within each utterance, the Environmental Sniffing framework did not allow noise transitions within an utterance. A noise classification rate of 82% was obtained. From this, a 9x9 error matrix ϵ was generated for use in digit recognition.

7.2. Digit Recognition Results:

7.2.1. Development Phase:

Environmental condition specific acoustic models were trained and used during recognition tests. Matrix $\mathbf{W}$ was generated by testing different acoustic conditions using different acoustic models. By using the $\mathbf{W}$ matrix, we calculated the normalized cost matrix $\mathbf{C}$ using Eq. 5. Using Eq. 6, with the *a prior* noise probabilities,

$$\vec{\mathbf{a}} = \quad 0.05 \quad 0.15 \quad 0.15 \quad 0.15 \quad 0.15 \quad 0.15 \quad 0.15 \quad 0.05 \quad ,$$

CPR was calculated as 92.1%, and the expected performance deviation was found to be 0.22% when the Environmental Sniffer uses the knowledge that only one acoustic condition is present within each utterance.

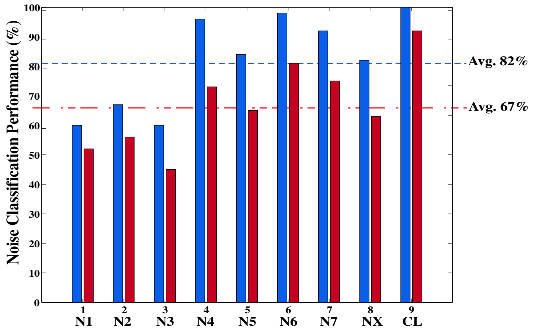

Fig. 4. Sniffing performance for each noise type (A) with(left bar in each pair)/(B) without(right bar in each pair) the prior knowledge that there was one environmental condition present in each utterance.

7.2.2. Test Phase:

Having established the environmental sniffer, and normalized cost matrix $\mathbf{C}$ for directing ASR model selection, we now turn to ASR system evaluation. We tested and compared the following 3 system configurations:

S1: Model matching was done using *a priori* knowledge of the acoustic noise condition.

S2: Model matching was done based on the environmental acoustic knowledge extracted from Environmental Sniffing.

S3: All acoustic condition dependent models were used in a parallel multi-recognizer structure (e.g. ROVER) without using any noise knowledge and the recognizer hypothesis with the highest path score was selected.

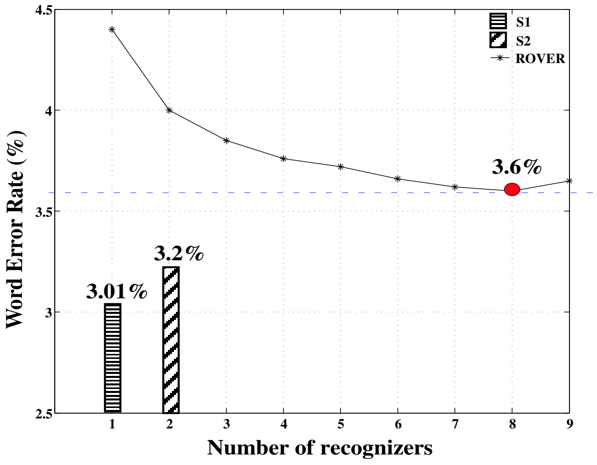

Fig. 5. Comparison of system configurations S1, S2, S3.

As shown if Fig. 5, system S1 achieved the lowest WER (i.e., 3.01%) since the models were matched perfectly to the acoustic condition during decoding.From the development phase, we know that the expected performance deviation was 0.22 for a model matching scheme employing Environmental Sniffing, which means that we can expect a WER value of 3.01+0.22=3.23% for S2. Experimentally, the WER for S2 was 3.2% using 2 CPU's (1 CPU for digit recognition, 1 CPU for sniffing acoustic conditions), which was close to the expected value of 3.23% (Note: in Fig. 5, we plot system S2 2 CPU even though only 1 ASR engine was used). S3 achieved a WER of 3.6% by using 8 CPU's. When we compare S2 and S3, we see that a relative 11.1% WER improvement was achieved, while requiring a relative 75% **reduction** in CPU resources. These results confirm the advantage of using Environmental Sniffing over a ROVER paradigm.

8. DISCUSSION

There are two critical points to consider when integrating Environmental Sniffing into a speech task. First, and the most important, is to set up a configuration such as S1 where prior noise knowledge can be fully used to yield the lowest WER. This will require understanding of the sources of errors and finding specific solutions assuming that there is prior acoustic knowledge. For example, knowing which speech enhancement scheme or model adaptation scheme is best for a specific acoustic condition is required.

Secondly, a reliable cost matrix should be provided to the Environmental Sniffing so the subsequent speech task can calculate the expected performance in making an informed adjustment in the trade-off between performance and computation. For our experiments, we considered evaluation results for Environmental Sniffing where it is employed to find the *highest* possible acoustic condition so that correct acoustic condition dependent model could be used. This is most appropriate for the goal of determining a single solution for the speech task problem at hand. If the expected performance for the system employing Environmental Sniffing is lower than the performance of a ROVER system, it may be useful to find the n most probable acoustic condition types among N acoustic conditions. In the worst case, the acoustic condition knowledge extracted from Environmental Sniffing could be ignored and the system will reduce to the traditional ROVER solution.

Finally, in an ASR task, in addition to determining the number of systems that should be combined, even a ROVER paradigm could take advantage of Environmental Sniffing to address open questions such as the combination order of the ASR system hypotheses, engage or disable preprocessing or normalization of system outputs prior to combination, etc. The goal therefore has been to emphasize that direct estimation of environmental conditions should provide important information to tailor a more effective solution to robust speech recognition systems.

We are currently working on evaluating our unsupervised noise analysis module within the existing framework. By the time of the conference, we will be able to present comparison between supervised noise analysis (human transcribed - segmented and clustered) and unsupervised noise analysis (automatically transcribed - segmented and clustered). Results for Environmental Sniffing vs ROVER in an in-vehicle recognition task will be presented by the time of the conference.

9. CONCLUSION

In this study, we have extended our previous proposed *Environmental Sniffing* [1, 2] framework by adding an unsupervised noise analysis module. We integrated the sniffing framework into an in-vehicle hands-free digit recognition system. The critical performance rate (CPR) was formulated for this task. The sniffing framework was compared to a ROVER solution in terms of WER and CPU usage in a model matching task where environmental condition dependent models were used during decoding. Results are presented by using supervised noise analysis module. In our experiments, the presented framework consistently outperformed the original ROVER solution by 11.1% in WER, while requiring 75% less CPU resources.

10. REFERENCES

[1] M. Akbacak, J.H.L. Hansen, "Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003.

[2] M. Akbacak, J.H.L. Hansen, "Environmental Sniffing: Robust Digit Recognition for an In-Vehicle Environment," *INTERSPEECH-2003/Eurospeech-2003*, pp.2177-2180, Geneva, Switzerland, September 2003.

[3] Y. Gong, "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*, v16-3, 261-91,1995.

[4] J. G. Fiscus, "A Post Processing System to yield reduced error rates: Recognizer Output Voting Error Reduction (ROVER)," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 347-54,1997.

[5] J. S. Lim, "Speech Enhancement", *Prentice Hall, Englewood Cliffs*, NJ, 1983.

[6] J.H.L. Hansen, "Speech Enhancement", Encyclopedia of Electrical and Electronics Engineering, John Wiley & Sons, vol. 20, pp. 159-175, 1999.

[7] R. Sarikaya, J.H.L. Hansen, "Improved Jacobian Adaptation for Fast Acoustic Model Adaptation for Noisy Speech Recognition", *Proc. of Intl. Conf. on Spoken Language Processing (ICSLP)*, vol. 3, pp. 702-705, October 2000.

[8] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, vol. 9, no. 2, pp. 171-185, April, 1995.

[9] M. Gales and S. Young, "Robust Continuous Speech Recognition using Parallel Model Combination", *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 352-359, September 1996.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990

[11] R. Sarikaya and J.H.L. Hansen, "High Resolution Speech Feature Parametrization for Monophone Based Stressed Speech Recognition", *IEEE Signal Processing Letters*, vol. 7, no. 7, pp. 182-185, July 2000.

[12] J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition", *Speech Communications, Special Issue on Speech Under Stress*, vol. 20(2), pp. 151-170, November 1996.

[13] S. Chen and P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127-132, February 1998

[14] B. Zhou and J.H.L. Hansen, "Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion", *Proc. of Inter. Conf. on Spoken Language Processing ICSLP-2000*, vol. 3, pp. 714-717, October 2000.

[15] J.H.L. Hansen, et al., "CU-Move: Analysis & Corpus Development for Interactive In-Vehicle Speech Systems," *Eurospeech*, v3, 2023-6, 2001.

[16] B. L. Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer," *Technical Report #TR-CSLR-2001-01*, 2001

[17] AT&T FSM Library,
http://www.research.att.com/sw/tools/fsm/description.html