ENVIRONMENTAL SNIFFING: NOISE KNOWLEDGE ESTIMATION FOR ROBUST SPEECH SYSTEMS

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 propose a framework for extracting knowledge concerning environmental noise from an input audio sequence and organizing this knowledge for use by other speech systems. To date, most approaches dealing with environmental noise in speech systems are based on assumptions concerning the noise, or differences in collecting and training on a specific noise condition, rather than exploring the nature of the noise. We are interested in constructing a new speech framework entitled *Environmental Sniffing* to detect, classify and track acoustic environmental conditions. The first goal of the framework is to seek out detailed information about the environmental characteristics instead of just detecting environmental changes. The second goal is to organize this knowledge in an effective manner to allow smart decisions to direct other speech systems. Our current framework uses a number of speech processing modules including the Teager Energy Operator (TEO) and a hybrid algorithm with T^2-BIC segmentation, noise language modeling and GMM classification in noise knowledge estimation. We define a new information criterion that incorporates impact of noise in Environmental Sniffing performance. We use an in-vehicle speech and noise environment as a test platform for our evaluations and investigate the integration of Environmental Sniffing into an Automatic Speech Recognition (ASR) engine in this environment. Noise classification experiments show that the hybrid algorithm achieves an error rate of 25.51 % , outperforming a baseline system by an absolute 7.08%.

1. INTRODUCTION

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

This problem has been taken into consideration especially in ASR applications since the recognition performance degrades substantially due to changes in the environment. One of the first ASR tasks that have changing environmental conditions is for automatic transcription of "Broadcast News" (BN). Several research groups have worked on this task to increase recognition performance. These studies [1, 2] have the underlying goal of training for acoustic conditions that are specific for each system (speech conditions include : F0- prepared, F1- spontaneous,F2- degraded acoustics, F3- music background, F4- noise background, F5- nonnative speakers, and FX- other speech) and directing the ASR engine to a single recognizer for each acoustic condition. The downside of this method is that it tries to model many different kinds of environmental conditions with a single model, with the hope that such a background noise model would be able to capture this huge variability.

Later, as computational power has increased with the help of high-speed computers, a parallel bank of recognizers has been used in a ROVER paradigm for tasks such as Speech In Noisy Environments (SPINE) where many different environmental conditions exist. Different recognizers intentionally employing a range of feature processing or adaptation methods are normal for a ROVER based LVCSR solution. This may involve different features during the feature extraction step, different noise compensation schemes in the enhancement step, or different model adaptation schemes individually or in parallel. Finally, the hypothesis with the highest probability at the output of the decoders is chosen as the final decision of the ROVER. Although significant improvement has been achieved using the ROVER paradigm, it is not optimal in terms of computational performance. It is also highly possible that one recognizer may not have the highest probability at all times during decoding, implying that the selected recognizer may be the best in a global sense but not in a local sense.

To overcome the disadvantages of these methods as well as to have acceptable error rates in ASR systems in changing environmental conditions, we propose a new speech framework called *Environmental Sniffing*. The goal will be to do smart tracking of environmental conditions and direct the ASR engine to use the best local solution specific to each environmental condition. For example, instead of running parallel feature extractors in a ROVER paradigm, the Environmental Sniffing framework will direct the ASR engine to use only one feature extractor which gives the best performance for a specific environmental condition. In this way, we optimize both the computational effort and overall system performance of the ASR.

On the other hand, Environmental Sniffing is also useful for automatic transcription of noise where the accuracy is much lower than that of transcription of speech. Considering the fact that there are no standards for noise transcription in audio material, it is critical to automatically transcribe environmental noise with high accuracy for more effective speech system training.

The organization of our paper is as follows. In Section 2, a general system architecture for Environmental Sniffing is presented. In Section 3, we specialize the general framework for sniffing environmental noise for in-vehicle systems. In Section 4, evaluations of the framework integrated into an in-vehicle ASR engine is presented. Section 5 discusses some further research issues for sniffing. Conclusion is given in Section 6.

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

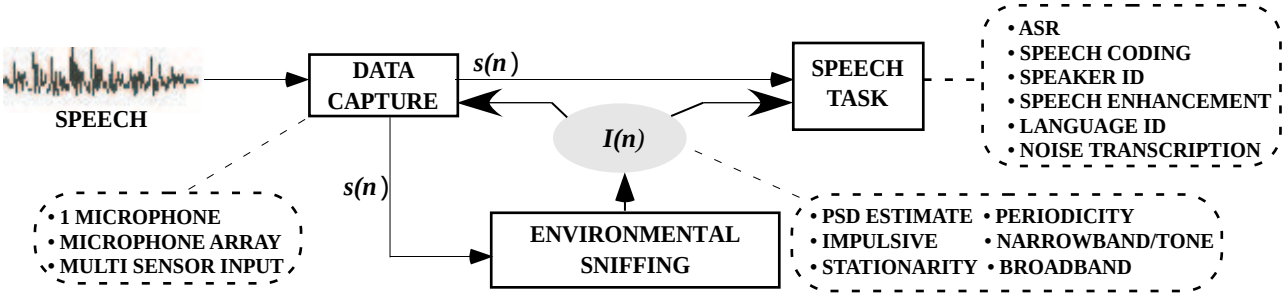

Fig. 1. Environmental Sniffing system architecture diagram.

2. SYSTEM ARCHITECTURE

Environmental Sniffing can be integrated into any speech task having some degree of concern about acoustic environmental conditions. Environmental Sniffing extracts knowledge about the acoustic environmental conditions and passes this knowledge to the speech task. A proposed general system architecture diagram 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 noise knowledge as $I(n)$ which is a function of $s(n)$. In a sample scenario, $s(n)$ may be the audio data recorded 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, $I(n)$ may also contain a suggestion to use one of several adaptation schemes (Jacobian adaptation, MLLR, PMC, etc.) which gives the best performance for the environmental noise knowledge estimated through Environmental Sniffing.

3. IN-VEHICLE ENVIRONMENTAL SNIFFING

Within the framework of Environmental Sniffing from Fig. 1, we specialize our solution for an in-vehicle hands-free car navigation environment. 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 [3].

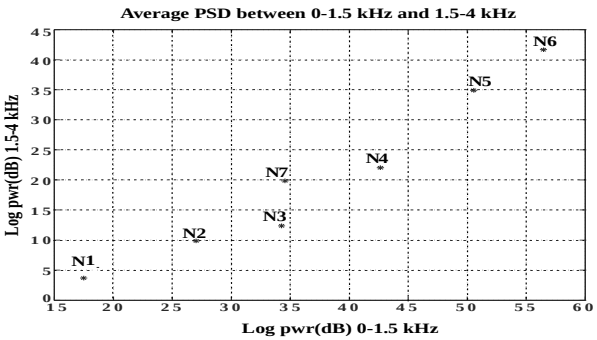

Fig. 2. Scatter plot of low (0-1.5 kHz) versus high (1.5-4 kHz) frequency noise dB-energy content for noises N1 through N7.

Having collected in-vehicle acoustic data (i.e., in a Blazer SUV) using a 17 mile route which contains samples of all driving conditions expected for use in city and rural areas, we identified the primary noise conditions of interest (noise conditions include: N1- idle noise consisting of the engine running with no movement and windows closed, N2- city driving without traffic and windows closed, N3- city driving with traffic and windows closed, N4- highway driving with windows closed, N5- highway driving with windows 2 inches open, N6- highway driving with windows half-way down, N7- windows 2 inches open in city traffic, N0- others), which are considered as long term acoustic environmental conditions. Other acoustic conditions (idle position with air-conditioning on, etc.) are matched to these primary classes having the closest acoustic characteristic. Fig. 2 shows the average power spectrum density for low (0-1.5 kHz) versus high (1.5-4 kHz) frequency energy content of long term noises. The diversity of noise energy content suggests that a single noise model would not be capable of addressing changing noise conditions for a subsequent speech task.

Short term acoustic environmental conditions occurring within long term conditions include TS- turn signal noise, WB- wiper blade noise, TN- tone noise, IM- impulsive noise. These conditions are expected to be present in conjunction with one of the long-term noises.

As shown in Fig. 3, a hybrid method of T^2-BIC segmentation and GMM classification followed by a decision smoothing is used to detect, classify and track long-term noises. T^2-BIC uses Hotelling's T^2-Statistic to pre-rank potential acoustic break points which are evaluated using a Bayesian Information Criterion [4].

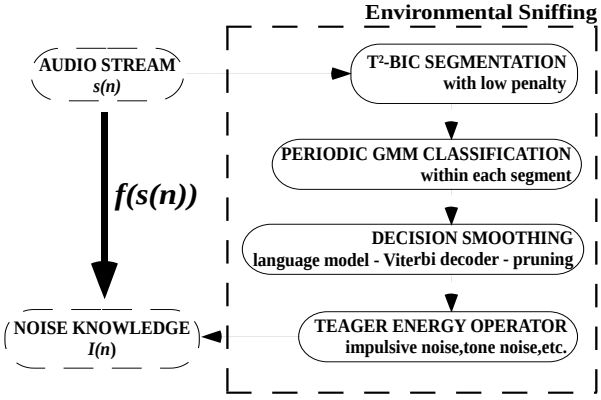

Fig. 3. Flow diagram for Environmental Noise Sniffing.

As Fig. 3 shows, the incoming audio stream is first segmented into acoustically homogeneous speech blocks using our T^2-BIC 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). Within each segment, GMM classification runs periodically to classify each non-overlapping T-frame-length block.

Decision smoothing is applied to the resulting decision sequence of each segment. This process is similar to Language Modeling, considering the fact that some noise transitions are not possible although they may appear at the output of the GMM classifier. Transition probabilities are generated from training data using bi-gram language modeling with a noise type for each 15-frame word block. Calculated transition probabilities are shown in Fig. 4.

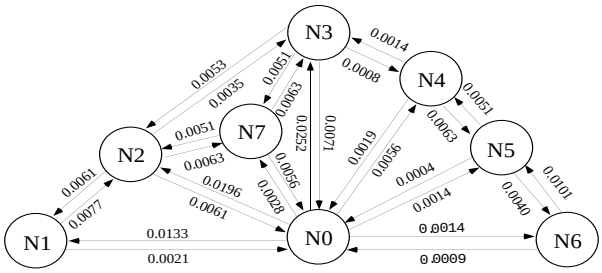

Fig. 4. Noise Language Modeling.

We use Viterbi decoding to find the most likely decision sequence given the classification probability list of each decision region within the segment. Each noise class has an initial probability which is proportional to the number of occurrences within the N-best position at the classifier output along the segment. Noise classes whose number of occurrences within the N-best position is less than a threshold are pruned during decision smoothing. We can formulate this as follows:

$$\alpha_1 n_1 + \alpha_2 n_2 + \ldots + (\alpha)_N n_N \leq \gamma \qquad (1)$$

where n_i: occurrence number in the i^{th} position of the score list, α_i: corresponding weight coefficient, and γ: threshold.

Since our Environmental Sniffing framework is not a speech system itself and works with other speech systems, noise knowledge detection performance for each noise type (P_i) should be weighted by a coefficient which is determined by the importance that noise type plays in the speech application with Environmental Sniffing (i.e., if noise impacts the speech task performance significantly, impact coefficient I is set high). For in-vehicle ASR, these coefficients ($I_1, I_2, \ldots, I_n$) will reflect the impact each noise type has on WER. We can formulate this as follows:

$$Critical\ performance\ rate \triangleq \sum_{i=1}^{n} I_i P_i \quad \sum_{i=1}^{n} I_i = 1 \qquad (2)$$

With this performance rate measure, the potential output score can range from 0-100 if P_i is a classification rate, or 0-1 if P_i is a probability.

4. EVALUATIONS

We evaluate the performance of our framework using an in-vehicle noise database of 3 hours collected in 6 experimental runs using the same route and the same vehicle on different days and hours. A microphone array and 8-channel digital recorder previously used for CU-Move in-vehicle speech data collection were employed [3]. The database does not contain speech. Fifteen noise classes are transcribed during the data collection by a transcriber sitting in the car. The time tags are generated instantly by the transcriber. After data collection, some noise conditions are grouped together, resulting in 8 acoustically distinguishable noise classes as listed in Sec. 3. For each noise class, a 4-mixture GMM is trained using 2.5 hours of data. We use 12 dimensional MFCC feature vectors during our evaluations. In both training and test data, long-term and short-term noise conditions are approximately equally balanced across time.

4.1. Long Term Noise

First, we test long-term noise classification error performance by running the classifier periodically with a period of 15 frames without segmenting the test data. Fig. 5 shows noise classification error performance by selecting the most likely model (solid bar to left in each pair) [avg. 34.73% error], and using the two highest probable models (cross-hatch bar to right in each pair) [avg. 13.23% error]. Some noise types (N4-highway driving, windows closed) are significantly affected by selecting the top two models out of 8 in the noise space.

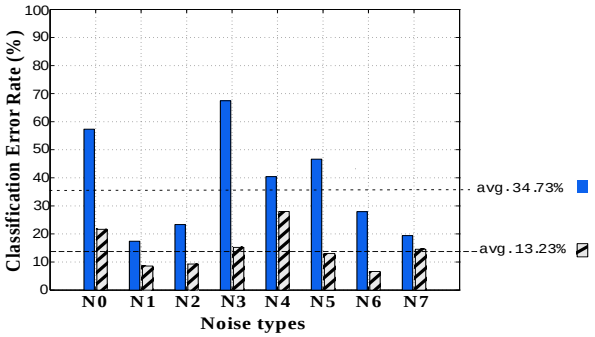

Fig. 5. Classification error performance of having the correct noise type in the first position (1^{st} bar in each set) and first two positions (2^{nd} bar in each set).

In our "Classical classification algorithm" for noise, a segment of data is scored once. As shown in Fig. 3, our "Hybrid Algorithm" has periodical classifications within a segment and subsequently smoothes the final decision sequence using the language model and pruning.

Next, we segment the noise test data using T^2-BIC with different false alarm penalties ($\lambda = \{0.3, 0.4, 0.5, 0.6\}$). During decision smoothing in the hybrid algorithm, we use the values $N = 2$, $\alpha_1 = 0.7$, $\alpha_2 = 0.3$, and pruning threshold $\gamma = 0.7$. Fig. 6 shows error rates for both methods. You can see that classical method is worse than the hybrid algorithm in terms of classification performance even if the hand label segmentation is provided.

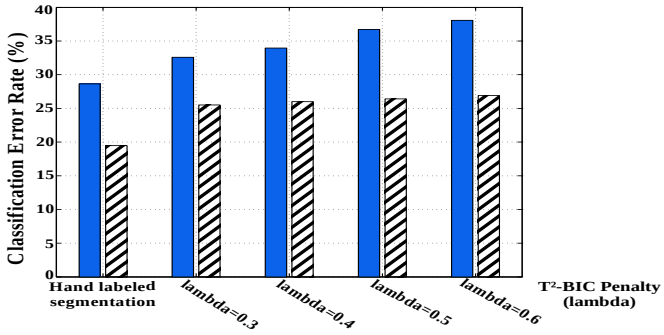

Fig. 6. Error performance of the classical method (1^{st} bar in each set) and the hybrid algorithm (2^{nd} bar in each set) with hand label segmentation, and a range of lambda values for T^2-BIC segmentation.

To calculate the overall performance using Eqn. 2, we ran speech recognition tests using CSLR's Large Vocabulary Continuous Speech Recognizer SONIC [5] on the TI-DIGITS database after degrading the clean speech with our noise types at 10 dB SNR. Models trained from clean speech were used for testing. WER results are shown in Table 1 as well as the impact I-measures of each noise type.

Degrading noise	N01	N02	N03	N04	N05	N06	N07
WER	1.1%	2.3%	2.7%	4.1%	8.1%	8.5%	3.7%
I-measure	0.04	0.08	0.09	0.13	0.27	0.28	0.11

Table 1. Speech Recognition Tests.

I_i's are assigned proportionally to WER's and they sum to one. Using Eqn. 2 with these values, we found the critical performance rate to be 65.41% for the classical classification method and 69.61% for the hybrid algorithm using 0.3 as the penalty parameter for T^2-BIC.

4.2. Discussion on Detecting Short Term Noise

We have the following assumptions about the human auditory system: hearing is the process of detecting energy at a particular frequency and the human auditory system is assumed to be a filtering process which partitions the entire audible frequency range into many critical bands. These assumptions provide motivation for use of the Teager Energy Operator (TEO) [6], to detect impulsive noise, tone noise and periodic noise observed in the in-vehicle environment since they appear as sudden energy changes and occupy a certain frequency band. What distinguishes these energy changes from those appearing during speech is that they do not have an observed modulation scheme like speech. Using this knowledge, we can automatically detect short-term noises within noisy-speech. Fig. 7 gives an idea of how TEO processing works for turn signal which occupies narrow time slots and a wide frequency band. The last figure (Fig. 7-c) clearly shows detection locations where turn signal noise is present.

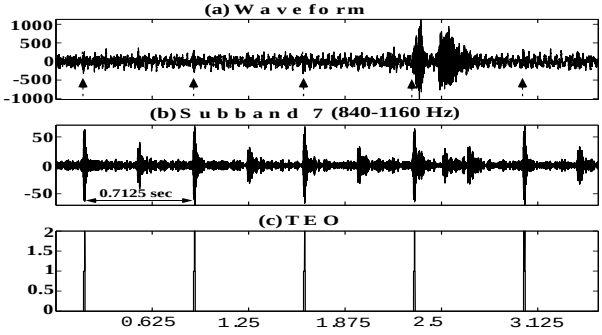

Fig. 7. Applying TEO processing to the environmental condition where the turn signal is on and the long-term noise is city driving with traffic, windows closed (N3).

5. DISCUSSION

The main goal of Environmental Sniffing is to extract knowledge about environmental noise that exists within continuous speech. Clearly, many studies have considered ASR in noise [7], and we believe that environmental sniffing will offer attractive alternatives to improve ASR performance (knowledge to improve features or model adaptation, reduce ROVER requirements). As a first step towards this goal, in our evaluations, we focused on extracting knowledge about the acoustic environmental noise using a noise-only audio database. However, while constructing the framework, we provide sufficient flexibility to easily move towards a subsequent step and to allow the same framework to be used for noisy-speech sections in audio streams as well. We are presently working on a broad class monophone recognition based framework to extract environmental noise knowledge from an audio stream consisting of both noisy-silence and noisy-speech (e.g., similar to our speech activity detection [SAD] work previously reported [8]). After defining a set of broad phone classes (e.g. nasals, unvoiced fricatives, voiced fricatives, etc.), we can generate monophone model sets where each corresponds to a noise type by degrading the clean monophone models with noise. In addition to these models, a silence model will also be included for each noise type. If we use 10 broad class monophones, we will have 10 clean monophone models, 10xN noisy monophone models, 1 clean silence model and N noisy-silence models, for a total of $(N + 1)$x11 models. Due to the pruning method used in the existing framework, the increase in search space will be less than a linear increase when we have more noise types. It will also be straightforward to use language modeling to calculate the transition probabilities from one monophone model set to another.

Another important issue is handling new in-coming noises within the framework, in other words, adapting Environmental Sniffing to new environmental noise types. Since there is a garbage noise model (N0) within the existing framework, we can keep track of the data classified as N0 and cluster to check if there is a sufficient data cluster to train a new additional noise model. We can also use the previous classification results to check how much the new model differs from existing ones by comparing the score distribution of the new model with existing ones.

6. CONCLUSION

In this paper, we have addressed the problem of changing acoustic environmental conditions in speech tasks. We proposed a new framework entitled *Environmental Sniffing* to detect, classify and, track changing acoustic environmental conditions and extract knowledge about the environmental noise. After proposing a general framework, we specialized the sniffer to an in-vehicle speech application. Novel aspects included a number of knowledge based processing steps such as T^2-BIC segmentation, noise language modeling, GMM classification and TEO processing. We believe such processing will provide significant knowledge to subsequent speech processing tasks and thereby increase robust speech performance.

7. REFERENCES

[1] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," *Speech Recognition Workshop (DARPA)*, Feb. 1997.

[2] R. Bakis, S. Chen, et al., "Transcription of BN Shows with the IBM LVCSR System," *Speech Recog. Workshop (DARPA)*, 1997.

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

[4] B. Zhou, J.H.L. Hansen, "Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion," *Proc. ICSLP*, v3, 714-7, 2000.

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

[6] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress," *IEEE Trans. on Speech & Audio Processing*, v9-2, 201-16, March 2001.

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

[8] R. Sarikaya, J.H.L. Hansen, "Robust detection of Speech Activity in the Presence of Noise," *Proc. ICSLP*, v4, 1455-8, 1998.