Chapter 2

CU-MOVE: ADVANCED IN-VEHICLE SPEECH SYSTEMS FOR ROUTE NAVIGATION[1]

John H.L. Hansen, Xianxian Zhang, Murat Akbacak, Umit H. Yapanel, Bryan Pellom, Wayne Ward, Pongtep Angkititrakul

Robust Speech Processing Group, Center for Spoken Language Research, University of Colorado at Boulder, Boulder, Colorado 80309-0594, USA
Email: John.Hansen@colorado. edu

Abstract: In this chapter, we present our recent advances in the formulation and development of an in-vehicle hands-free route navigation system. The system is comprised of a multi-microphone array processing front-end, environmental sniffer (for noise analysis), robust speech recognition system, and dialog manager and information servers. We also present our recently completed speech corpus for in-vehicle interactive speech systems for route planning and navigation. The corpus consists of five domains which include: digit strings, route navigation expressions, street and location sentences, phonetically balanced sentences, and a route navigation dialog in a human Wizard-of-Oz like scenario. A total of 500 speakers were collected from across the United States of America during a six month period from April-Sept. 2001. While previous attempts at in-vehicle speech systems have generally focused on isolated command words to set radio frequencies, temperature control, etc., the CU-Move system is focused on natural conversational interaction between the user and in-vehicle system. After presenting our proposed in-vehicle speech system, we consider advances in multi-channel array processing, environmental noise

[1] This work was supported in part by DARPA through SPAWAR under Grant No. N66001-002-8906, from SPAWAR under Grant No. N66001-03-1-8905, in part by NSF under Cooperative Agreement No. IIS-9817485, and in part by CSLR Center Member support from Motorola, HRL, Toyota CR&D, and CSLR Corpus Member support from SpeechWorks, Infinitive Speech Systems (Visteon Corp.), Mitsubishi Electric Research Lab, Panasonic Speech Technology Lab, and VoiceSignal Technologies.

sniffing and tracking, new and more robust acoustic front-end representations and built-in speaker normalization for robust ASR, and our back-end dialog navigation information retrieval sub-system connected to the WWW. Results are presented in each sub-section with a discussion at the end of the chapter.

Keywords: Automatic speech recognition, robustness, microphone array processing, multi-modal, speech enhancement, environmental sniffing, PMVDR features, dialog, mobile, route navigation, in-vehicle

1. INTRODUCTION: HANDS-FREE SPEECH RECOGNITION/DIALOG IN CARS

There has been significant interest in the development of effective dialog systems in diverse environmental conditions. One application which has received much attention is for hands-free dialog systems in cars to allow the driver to stay focused on operating the vehicle while either speaking via cellular communications, command and control of vehicle functions (i.e., adjust radio, temperature controls, etc.), or accessing information via wireless connection (i.e., listening to voice mail, voice dialog for route navigation and planning). Today, many web based voice portals exist for managing call center and voice tasks. Also, a number of spoken document retrieval systems are available for information access to recent broadcast news content including SpeechBot by HP-Compaq)[30] and the SpeechFind for historical digital library audio content (RSPG-CSLR, Univ. Colorado)[29]. Access to audio content via wireless connections is desirable in both commercial vehicle environments (i.e., obtaining information on weather, driving conditions, business locations, etc.), points of interest and historical content (i.e., obtaining audio recordings which provide a narrative of historical places for vacations, etc.), as well as in military environments (i.e., information access for coordinating peacekeeping groups, etc.).

This chapter presents our recent activity in the formulation of a new in-vehicle interactive system for route planning and navigation. The system employs a number of speech processing sub-systems previously developed for the DARPA CU Communicator[1] (i.e., natural language parser, speech recognition, confidence measurement, text-to-speech synthesis, dialog manager, natural language generation, audio server). The proposed CU-Move

system is an in-vehicle, naturally spoken mixed initiative dialog system to obtain real-time navigation and route planning information using GPS and information retrieval from the WWW. A proto-type in-vehicle platform was developed for speech corpora collection and system development. This includes the development of robust data collection and front-end processing for recognition model training and adaptation, as well as a back-end information server to obtain interactive automobile route planning information from WWW.

The novel aspects presented in this chapter include the formulation of a new microphone array and multi-channel noise suppression front-end, environmental (sniffer) classification for changing in-vehicle noise conditions, and a back-end navigation information retrieval task. We also discuss aspects of corpus development. Most multi-channel data acquisition algorithms focus merely on standard delay-and-sum beamforming methods. The new noise robust speech processing system uses a five-channel array with a constrained switched adaptive beamformer for the speech and a second for the noise. The speech adaptive beamformer and noise adaptive beamformer work together to suppress interference prior to the speech recognition task. The processing employed is capable of improving SegSNR performance by more than 10dB, and thereby suppress background noise sources inside the car environment (e.g., road noise from passing cars, wind noise from open windows, turn signals, air conditioning noise, etc.).

This chapter is organized as follows. In Sec. 2, we present our proposed in-vehicle system. In Sec. 3, we discuss the CU-Move corpus. In Sec. 4, we consider advances in array processing, followed by environmental sniffing, and automatic speech recognition (ASR), and our dialog system with connections to WWW. Sec. 5 concludes with a summary and discussion of areas for future work.

2. CU-MOVE SYSTEM FORMULATION

The problem of voice dialog within vehicle environments offers some important speech research challenges. Speech recognition in car environments is in general fragile, with word-error-rates (WER) ranging from 30-65% depending on driving conditions. These changing environmental conditions

include speaker changes (task stress, emotion, Lombard effect, etc.)[16,31] as well as the acoustic environment (road/wind noise from windows, air conditioning, engine noise, exterior traffic, etc.).

Recent approaches to speech recognition in car environments have included combinations of basic HMM recognizers with front-end noise suppression[2,4], environmental noise adaptation, and multi-channel concepts. Many early approaches to speech recognition in the car focused on isolated commands. One study considered a command word scenario in car environments where an HMM was compared to a hidden Neural Network based recognizer[5]. Another method showed an improvement in computational requirements with front-end signal-subspace enhancement used a DCT in place of a KLT to better map speech features, with recognition rates increasing by 3-5% depending on driving conditions[6]. Another study[7] considered experiments to determine the impact of mismatch between recognizer training and testing using clean data, clean data with car noise added, and actual noisy car data. The results showed that starting with simulated noisy environment train models, about twice as much adaptation material is needed compared with starting with clean reference models. The work was later extended[8] to consider unsupervised online adaptation using previously formulated MLLR and MAP techniques. Endpoint detection of phrases for speech recognition in car environments has also been considered[9]. Preliminary speech/noise detection with front-end speech enhancement methods as noise suppression front-ends for robust speech recognition have also shown promise[2,4,10,11]. Recent work has also been devoted to speech data collection in car environments including SpeechDat.Car[12], and others [13]. These data concentrate primarily on isolated command words, city names, digits, etc. and typically do not include spontaneous speech for truly interactive dialogue systems. While speech recognition efforts in car environments generally focus on isolated word systems for command and control, there has been some work on developing more spontaneous speech based systems for car navigation [14,15], however these studies use a head-worn and ceiling mounted microphones for speech collection and limit the degree of naturalness (i.e., level of scripting) for navigation information exchange.

In developing CU-Move, there are a number of research challenges which must be addressed to achieve reliable and natural voice interaction within the car environment. Since the speaker is performing a task (driving the vehicle), a measured level of user task stress will be experienced by the driver and

therefore this should be included in the speaker modeling phase. Previous studies have clearly shown that the effects of speaker stress and Lombard effect (i.e., speaking in noise) can cause speech recognition systems to fail rapidly[16]. In addition, microphone type and placement for in-vehicle speech collection can impact the level of acoustic background noise and ultimately speech recognition performance. Figure 2-1 shows a flow diagram of the proposed CU-Move system. The system consists of front-end speech collection/processing tasks that feed into the speech recognizer. The speech recognizer is an integral part of the dialogue system (tasks for Understanding, Discourse, Dialogue Management, Text Generation, and TTS). An image of the microphone used in the array construction is also shown (Figure 2-2). The back-end processing consists of the information server, route database, route planner, and interface with the navigation database and navigation guidance systems. Here, we focus on our efforts in multi-channel noise suppression, automatic environmental characterization, robust speech recognition, and a proto-type navigation dialogue.

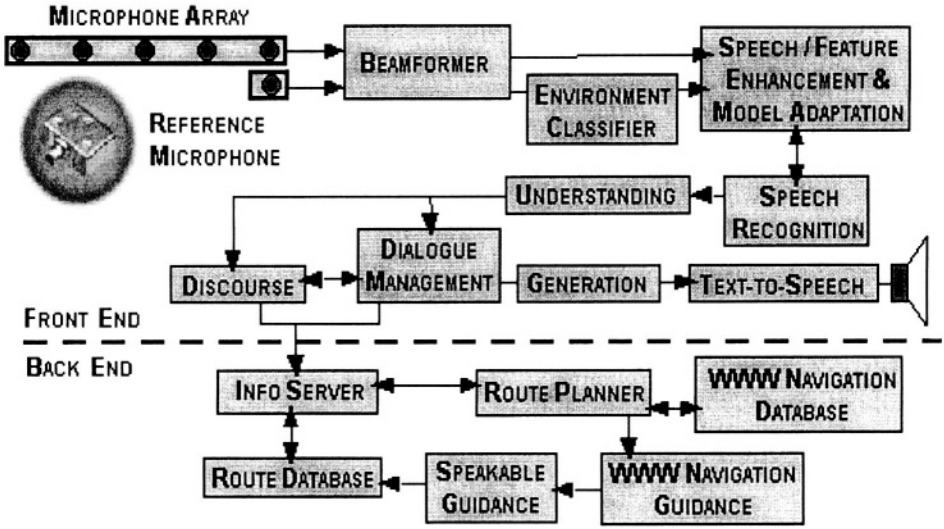

Figure 2-1. Flow Diagram of CU-Move Interactive Dialogue System for In-Vehicle Route Navigation

3. CU-MOVE CORPUS DEVELOPMENT

As part of the CU-Move system formulation, a two phase data collection plan was developed. Phase I focused on collecting acoustic noise and probe speech from a variety of cars and driving conditions. The outcome of Phase I was to determine the range of noise conditions across vehicles, and select one vehicle for Phase II collection that is representative of the typical noise domains experienced while driving. Eight vehicles were used in Phase I analysis (e.g., compact and two mid-size cars, small and medium pickup trucks, passenger van, sport utility vehicle (SUV), cargo van). We considered 14 noise conditions in actually driving scenarios. Figure 2-2 summarizes some of the results obtained from the study, with further details presented in [26]. The noise level was highest with windows open 2 inches traveling 65mph on the highway, and most quiet when the car was idle at a stop light. After detailed analysis, we determined the SUV represented the mid-range noise conditions (noise levels were high for compact cars and low for pickup trucks).

Next, Phase II speech collection was performed. Since the speaker is experiencing some level of stress by performing the task of driving the vehicle, this should be included in the speaker modeling phase. While Lombard effect can be employed, local state and federal laws in the United States limit the ability to allow subjects in this data collection to operate the vehicle and read prompts from a display. We therefore have subjects seated in the passenger seat, with prompts given on a small flat panel display attached to the dashboard to encourage subjects to stay focused on the roadway ahead. Speech data collection was performed across 6 U.S. cities that reflect regional dialects. These cities were selected to be mid-size cities, in order to increase the prospects of obtaining subjects who are native to that region. A balance across gender and age brackets was also maintained. The driver performed a fixed route similar to what was done for Phase I data collection so that a complete combination of driving conditions (city, highway, traffic noise, etc.) was included. The format of the data collection consists of five domains with four *Structured Text Prompt* sections and one *Wizard-of-Oz* (WOZ) dialog section:

Navigation Phrases: collection of phrases useful for In-Vehicle navigation interaction [prompts are fixed for all speakers]. Examples include: “Where is the closest gas station?” “How do I get to 1352 Pine Street?” “Which exit do I

take?" "Is it a right or left turn?" "How do I get to the airport?" "I'm lost. Help me."

Digit Sequences: each speaker produced 16 digit strings from a randomized 75 digital string set. Examples include: telephone numbers (425-952-54o0), random credit card numbers (1234-5621-1253-5981), and individual numbers (0,o,#86, *551).

Say and Spell Addresses: a randomized set of 20 strings of words/addresses were produced, with street names spelled. Some street names are used for all cities, some were drawn from local city maps. Examples include: Park Place, Ivy Circle, 3215 Marine Street, 902 Olympic Boulevard.

Phonetically Balanced Sentences: each speaker produced a collection of between 20-30 phonetically balanced from a set of 2500 sentences [prompts are randomized]. Examples include: "This was easy for us." "Jane may earn more money by working hard."

Dialog Wizard - of - Oz Collection: each speaker from the field called an on-line navigation system at CSLR, where a human wizard-of-oz like (WOZ) operator would guide the caller through three different navigation routes determined for that city. More than 100 potential destinations were previously established for each city between the driver and WOZ human operator, where detailed route information was stored for the operator to refer to while the caller was on the in-vehicle cell-phone. The list of establishments for that city were points of interest, restaurants, major intersections, etc. (e.g., "How do I get to the closest police station?", "How do I get to the Hello Deli?"). The user calls using a modified cell-phone in the car, that allows for data collection using one of the digital channels from our recorder. The dialog was also recorded at CSLR where the WOZ operator was interacting with the field subject.

The 500 speaker corpus was fully transcribed, labeled, spell checked, beamformed/processed and organized for distribution. The un-processed version contains well over 600GB of data, and the processed version consists of a hard-disk release of approximately 200GB. Figure 2-3 shows the age distribution of the CU-Move corpus (further details presented in [26,27]).

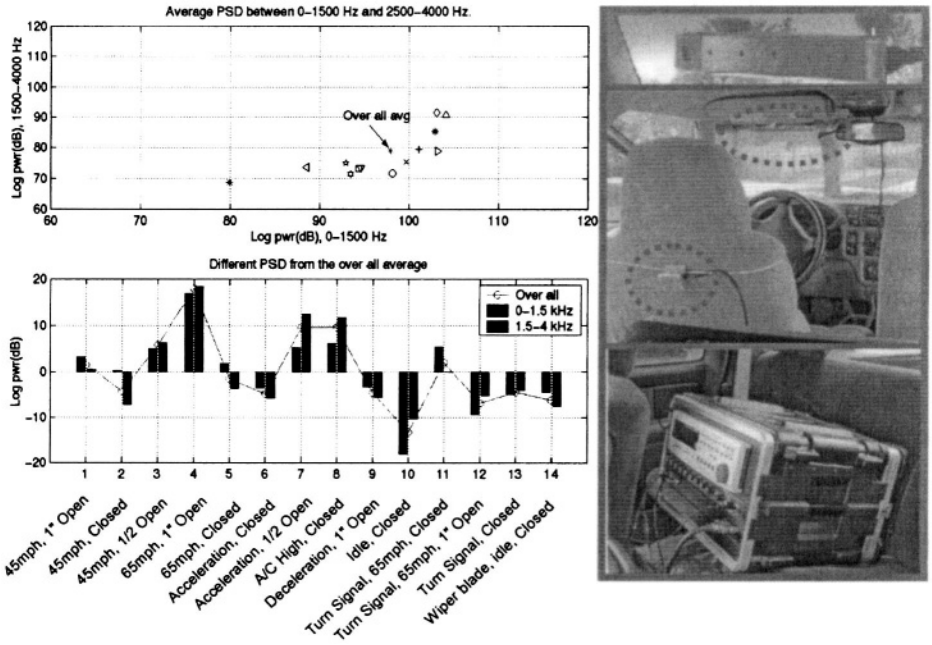

Figure 2-2. (a) Analysis of average power spectral density for low (0-1.5kHz) and high (1.5-4.0kHz) frequency bands for 14 car noise conditions from Phase-I data collection. Overall average noise level is also shown. (b) Photos show corpus collection setup: constructed microphone array (using Knowles microphones), array and reference microphone placement, constructed multi-channel DAT recorder (Fostex) with channel dependent level control and DC-to-AC converter.

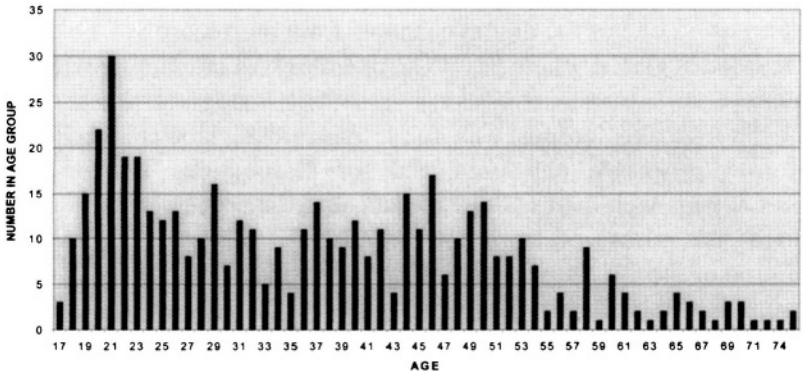

Figure 2-3. Age distribution (17-76 years old) of the 500 speaker CU-Move Corpus

In addition, a number of cross-transcriber reliability evaluations have been completed on the CU-Move corpus. Three transcribers were on the average, in agreement the majority of the time for parts 1-4 (prompts), with a 1.8% substitution rate when comparing transcriber hypotheses two at a time. When we consider the spontaneous route navigation WOZ part, transcriber files naturally had a higher difference, with a substitution rate of 3.0%. These numbers will depend on the clarity and articulation characteristics of the speakers across the six CU-Move dialect regions.

4. IN-VEHICLE SUB-SYSTEM FORMULATION

In this section, we discuss the formulation of our microphone array processing front-end, environmental sniffing, robust speech recognition system, and proto-type dialogue system.

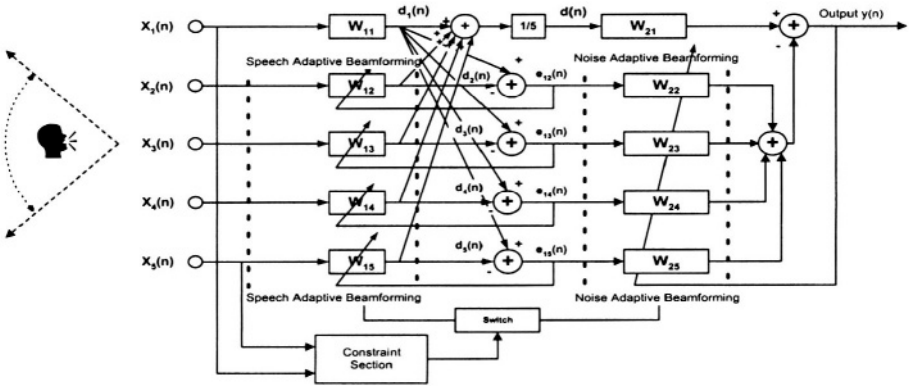

Figure 2-4. Flow diagram of the Proposed Constrained Switched Adaptive Beamforming (CSA -BF) algorithm.

4.1 Constrained Switched Adaptive Array-Processing (CSA-BF)

The proposed CSA-BF array processing algorithm consists of four parts: a constraint section (CS), a speech adaptive beamformer (SA-BF), a noise adaptive beamformer (NA-BF) and a switch. Figure 2-4 shows the detailed structure of CSA-BF, for a 5-microphone array. The CS is designed to identify potential speech and noise locations. If a speech source is detected, the switch will activate SA-BF to adjust the beam pattern and enhance the desired speech. At the same time, NA-BF is disabled to avoid speech leakage. If however, a noise source is detected, the switch will activate NA-BF to adjust the beam pattern for noise and switch off SA-BF processing to avoid the speech beam pattern from being altered by the noise. The combination of SA-BF and NA-BF processing results in a framework that achieves noise cancellation for interference in both time and spatial orientation. Next, we consider each processing stage of the proposed CSA-BF scheme.

4.1.1 Constraint Section

Many source localization methods have been considered in the past with effective performance for large microphone arrays in conference rooms or large auditoriums. Their ability to perform well in changing noisy car conditions has not been documented to the same degree, but is expected to be poor. Here, we propose three practical constraints that can be used to separate speech and noise sources with high accuracy.

- *Criterion 1 (Maximum averaged energy):* Since speech coming from the driver's direction will have on average the highest intensity of all sources present, therefore, we calculate the averaged signal TEO energy [18] frame by frame, and if this energy is greater than some threshold (Please refer to [19] for the details, we take the current signal frame as speech candidate.
- *Criterion 2 (LMS adaptive filter):* In order to separate the front-seat driver and passenger, we choose the adaptive LMS filter method and incorporate the geometric structure of the microphone array to locate the source.
- *Criterion 3 (Bump noise detector)* This final criterion is set to avoid instability in the filtering process which is affected by impulsive noise with high-energy content, such as road impulse/bump noise.

Finally, we note that the signal is labeled as speech if and only if all three criteria are satisfied.

4.1.2 Speech Adaptive Beamformer (SA-BF)

The function of SA-BF is to form an appropriate beam pattern to enhance the speech signal. Since adaptive filters are used to perform the beam steering, we can change beam pattern with a movement of the source. The degree of adaptation steering speed is decided by the convergence behavior of the adaptive filters. In our implementation, we select microphone 1 as the primary microphone, and build an adaptive filter between it and each of the other four microphones. These filters compensate for the different transfer functions between the speaker and the microphone array. A normalized LMS algorithm updates the filter coefficients only when the current signal is detected as speech. There are two kinds of output from the SA-BF: namely the enhanced speech $d(n)$ and noise signal $e_{1i}(n)$, which are given as follows,

$$d(n) = \frac{1}{5}\sum_{i=1}^{5} \mathbf{w}_{1i}^{T}(n)\mathbf{x}_{1i}(n) \tag{1}$$

$$e_{1i}(n) = \mathbf{w}_{11}^{T}(n)\mathbf{x}_{1}(n) - \mathbf{w}_{1i}^{T}(n)\mathbf{x}_{i}(n) \tag{2}$$

$$\mathbf{w}_{1i}(n+1) = \mathbf{w}_{1i}(n) + \frac{2\mu}{\mathbf{x}_{i}^{T}(n)\mathbf{x}_{i}(n)} e_{1i}(n)\mathbf{x}_{i}(n) \tag{3}$$

for channels i=2,3,4,5, where $\mathbf{w}_{11}(n)$ is a fixed filter.

4.1.3 Noise Adaptive Beamformer (NA-BF)

The NA-BF processor operates in a scheme like a multiple noise canceller, in which both the reference speech signal of the noise canceller and the speech free noise references are provided by the output of the SA-BF. Since the filter coefficients w_{2i} are updated only when the current signal is detected as noise, they form a beam that is directed towards the noise, thus the reason to name it a noise adaptive beamformer (NA-BF). The output response is given as,

$$y(n) = \mathbf{d}\ (n)\mathbf{w}_{21}^{T}(n) - \sum_{i=2}^{5} \mathbf{w}_{2i}^{T}(n)\mathbf{e}_{1i}(n) \tag{4}$$

$$\mathbf{w}_{2i}(n+1) = \mathbf{w}_{2i}(n) + \frac{2\mu}{\mathbf{e}_{1i}^{T}(n)\mathbf{e}_{1i}(n)} e_{1i}(n)d(n) \tag{5}$$

for microphone channels i=2,3,4,5.

4.1.4 Experimental Evaluation

In order to evaluate the performance of the CSA-BF algorithms in noisy car environments, we process all available speakers in Release 1.1a [21,26,27] of the CU-Move corpus using both CSA-BF and DASB algorithms, and compared the results. This release consists of 153 speakers, of which 117 were from the Minneapolis, MN area. We selected 67 of these speakers that include 28 males and 39 females, which reflects 8 hours of data. In order to compare the result of CSA-BF with that of DASB thoroughly, we also investigated the enhanced speech output from SA-BF. For evaluation, we consider two different performance measures using CU-Move data. One measure is the Segmental Signal-to-Noise Ratio (SegSNR) [22] which represents a noise reduction criterion for voice communications. The second performance measure is Word Error Rate (WER) reduction, which reflects benefits for speech recognition applications. The Sonic Recognizer [23,25] is used to investigate speech recognition performance. During the recognizer evaluation, we used 49 speakers (23 male, 26 female) as the training set, and 18 speakers (13 male, 5 female) as the test set.

Table 2-1 summarizes average SegSNR improvement, average WER, CORR (word correct rate), SUB (Word Substitution Rate), DEL (Word Deletion Rate) and INS (Word Insertion Rate). Here, the task was on the digits portion of CU-Move corpus (further details are presented in [19]). Figure 2-5 illustrates average SegSNR improvement and WER speech recognition performance results. The average SegSNR results are indicated by the bars using the left-side vertical scale (dB), and the WER improvement is the solid line using the right-side scale (%).

Measure \ Method	chan3	DASB	SA-BF	CSA-BF
Ave. (dB) SegSNR	9.35	10.24	10.51	14.79
WER (%)	14.8	11.9	12	11
SUB (%)	7.9	6.8	6.6	6.2
DEL (%)	4.3	2.5	2.8	2.5
INS (%)	2.5	2.6	2.5	2.4
CORR (%)	87.7	90.7	90.5	91.3

Table 2-1. Average SegSNR (segmental signal-to-noise ratio), WER (word-error-rate), CORR (word correct rate), SUB (word substitution rate), DEL (word deletion rate) and INS (word insertion rate) for Reference Channel 3 Microphone (chan3) and three Array/Beamforming Scenarios: DASB (delay-and-sum beamforming), SA-BF (speech adaptive beamforming), CSA-BF (constrained switched adaptive beamforming).

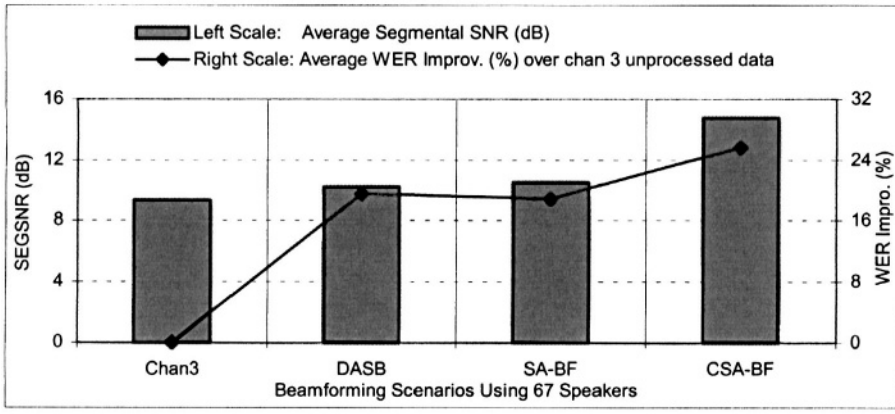

Figure 2-5. SEGSNR and WER Results for Reference Channel 3 Microphone (Chan3) and Array Processing/Beamforming Scenarios using 67 speakers from the CU-Move Corpus. Bar graph represents SegSNR in dB (using the left side scale), and line plot represents Avg. word error rate improvement (in %) (using the right side scale).

From these results (Table 2-1, Figure 2-5), we draw the following points:

1. Employing delay-and-sum beamforming (DASB) or the proposed speech adaptive beamforming (SA-BF), increases SegSNR slightly, but some variability exists across speakers. These two methods are able to improve WER for speech recognition by more than 19%.

2. There is a measurable increase in SegSNR and a decrease in WER when noise cancellation processing is activated (CSA-BF). With CSA-BF, SegSNR improvement is +5.5dB on the average, and also provides a relative WER improvement of 26%.

4.2 Environmental Sniffing

In this section we discuss our novel 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 which we have entitled *Environmental Sniffing* to detect, classify and track acoustic environmental conditions in the car environment (Figure 2-6, see [24,32]). 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 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 broad class monophone recognition in noise knowledge estimation. We define a new information criterion, *Critical Performance Rate* (CPR), that incorporates the impact of noise into Environmental Sniffing performance by weighting the rate of each error type with a normalized cost function. 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.

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. 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.

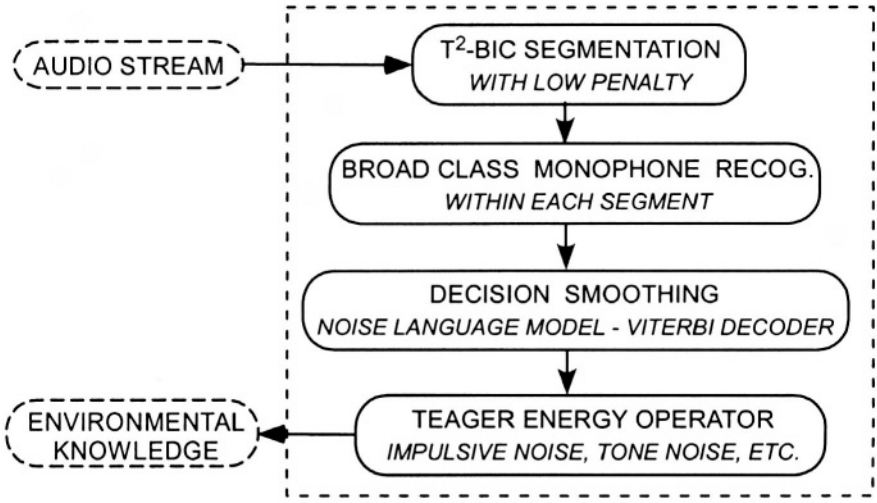

Figure 2-6. Flow Diagram for In-Vehicle Environmental Sniffing

We identified the following primary noise conditions of interest: (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, NX- 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.

Since the Environmental Sniffing framework is not a speech system itself, and must work with other speech systems, noise knowledge detection performance for each noise type should be calculated by weighting each

classification error type by a term which is conditioned on the importance that error type plays in the subsequent speech application employing Environmental Sniffing. In [32], 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. For this task, we can formulate the Critical Performance Rate as:

$$CPR = 1 - diag\{C \cdot \varepsilon^T\} \cdot \vec{a}^T, \quad (6)$$

where $\boldsymbol{\varepsilon^T}$ denotes the transposed error matrix for noise classification, and **C** is the normalized cost matrix. Since some noise conditions occur more frequently than others, each noise condition will have an *a priori* probability denoted as **a.** Each cost value is proportional with WER difference between the matched case and the mismatched case, which is the performance deviation of the ASR engine by using the wrong acoustic model during decoding instead of using the correct acoustic model. 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.

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 [24]. 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.

Using the sniffing framework presented in Figure 2-6, 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. Environmental condition specific acoustic models were trained and used during recognition tests. The Cost matrix **C** is calculated by testing different acoustic conditions using different

acoustic models. The overall critical performance rate (CPR from Eq. (6)) was calculated as 92.1%

Having established the environmental sniffer, and normalized cost matrix 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 (i.e., establish theoretical best performance – matched noise conditions), 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.

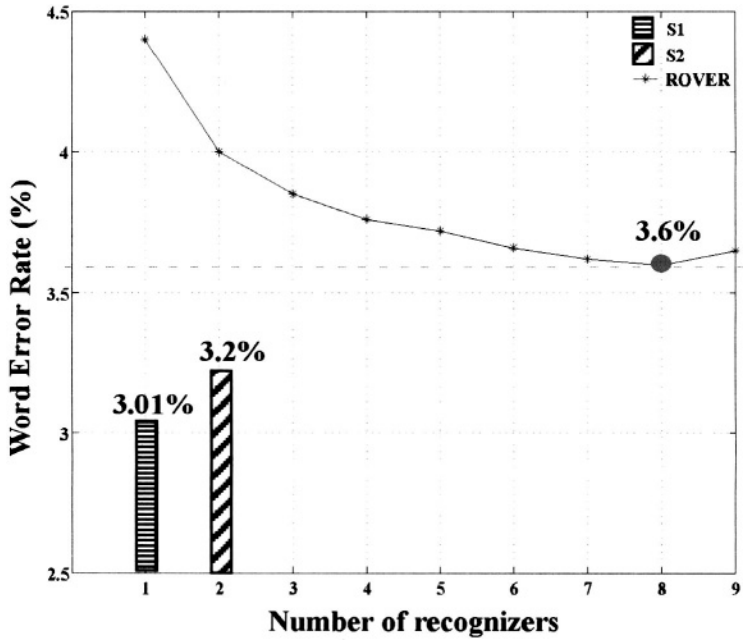

Figure 2-7. Word Error Rates for Digit Recognition Tests: S1 – matched noise model case, S2 – environmental sniffing model selection (1 CPU for sniffing, 1 CPU for ASR), S3 (ROVER) – employs up to 9 recognizers (i.e., CPUs) trained for each noise condition with ROVER selection.

As Figure 2-7 shows, system S1 achieved the lowest WER (i.e., 3.01%) since the models were matched perfectly to the acoustic condition during decoding. 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 Figure 2-7, we plot system S2 with 2 CPU's 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 an ASR ROVER paradigm.

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 (i.e., matched noise scenario). This will require an 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 the correct acoustic 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. The goal therefore in this section 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.

4.3 Robust Speech Recognition

The CU-Move system incorporates a number of advances in robust speech recognition including a new more robust acoustic feature representation and built-in speaker normalization. Here, we report results from evaluations using CU-Move Release 1.1 A data from the extended digits part aimed at phone dialing applications.

Capturing the *vocal tract transfer function* (VTTF) from the speech signal while eliminating other extraneous information, such as speaker dependent characteristics and pitch harmonics, is a key requirement for robust and accurate speech recognition [33, 34]. The vocal tract transfer function is mainly encoded in the *short-term spectral envelope* [35]. Traditional MFCCs use the *gross spectrum* obtained as the output of a non-linearly spaced filterbank to represent the spectral envelope. While this approach is good for unvoiced sounds, there is a substantial mismatch for voiced and mixed sounds [34]. For voiced speech, the formant frequencies are biased towards strong harmonics and their bandwidths are misestimated [34,35]. MFCCs are known to be fragile in noisy conditions, requiring additional compensation for acceptable performance in realistic environments [45,28].

Minimum Variance Distortionless Response (MVDR) spectrum has a long history in signal processing but recently applied successfully to speech modeling [36]. It has many desired characteristics for a spectral envelope estimation method, most important being the fact it estimates the spectral powers accurately at the *perceptually important harmonics,* thereby providing an *upper envelope* which has strong implications for robustness in additive noise. Since the upper envelope relies on the high-energy portions of the spectrum, it will not be affected substantially by additive noise. Therefore, using MVDR for spectral envelope estimation for robust speech recognition is feasible and useful [37].

4.3.1 MVDR Spectral Envelope Estimation:

For details of MVDR spectrum estimation and its previous uses for speech parameterization, we refer the reader to [36,37,38,39,40]. In the MVDR spectrum estimation, the signal power at a frequency, ω_l, is determined by filtering the signal by a specially designed FIR filter, h(n), and measuring the power at its output. The FIR filter, h(n), is designed to minimize its output

power subject to the constraint that its response at the frequency of interest, ω_l, has unity gain. This constrained optimization is a key aspect of the MVDR method that allows it to provide a lower bias with a smaller filter length than the Periodogram method [41]. The Mth order MVDR spectrum can be parametrically written as;

$$P_{MV}(\omega) = \frac{1}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}} = \frac{1}{\left|B(e^{j\omega})\right|^2} \tag{7}$$

The parameters, $\mu(k)$, can be obtained using the linear prediction (LP) coefficients, a_k, and the prediction error variance P_e [41].

$$\mu(k) = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{M-k} (M+1-k-2i) a_i a_{i+k}^*, & k = 0,\ldots,M \\ \mu^*(-k) & k = -M,\ldots,-1 \end{cases} \tag{8}$$

4.3.2 Direct Warping of FFT Spectrum

The aim of using a non-linearly spaced filterbank is to remove the harmonic information that exists in voiced speech and smooth out the spectrum. MVDR, on the other hand, can handle voiced speech by accurately modeling spectral powers at the perceptually important harmonics. Therefore, it is both *useful* and *safe* to remove the filterbank structure and incorporate the perceptual considerations by directly warping the FFT spectrum. The warping can be incorporated via a first order all pass system [42]. In fact, both Mel and Bark scales can be implemented by changing only one system parameter, α. We use the phase response of the first order system in Eq. (9) as the warping function given in Eq. (10),

$$H(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \tag{9}$$

$$\hat{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{10}$$

where α determines the degree of warping. For 16kHz sampled signals, α=0.42 and α=0.55 approximates the Mel and Bark scales, respectively.

4.3.3 PMVDR Algorithm

We can summarize the PMVDR algorithm as follows [37];

- *Step 1:* Obtain the perceptually warped FFT power spectrum,
- *Step 2:* Compute the "perceptual autocorrelations" by using IFFT on the warped spectrum,
- *Step 3:* Perform an Mth order LP analysis via Levinson-Durbin recursion using perceptual autocorrelation lags [41],
- *Step 4:* Calculate the Mth order MVDR spectrum using Eq. (7) from LP coefficients [36],
- *Step 5:* Obtain Cepstrum coefficients using the straightforward FFT-based approach [43].

A flow diagram for the PMVDR algorithm is given in Figure 2-8. The algorithm is integrated into the CU-Move recognizer as the *default acoustic feature front-end,* (further information and code can be obtained from the CU-Move web site [27]).

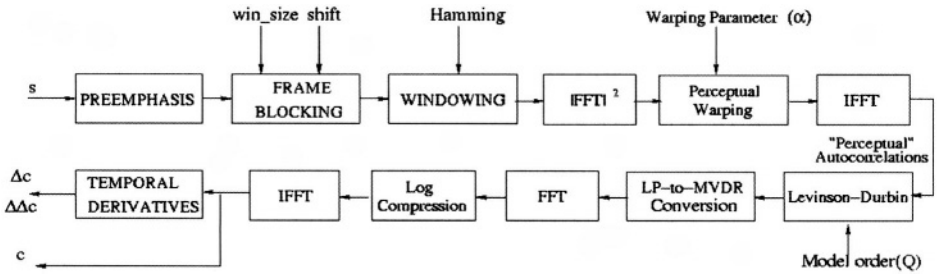

Figure 2-8. Flow Diagram of the PMVDR acoustic feature front-end

4.3.4 Experimental Evaluation

We evaluate the performance of PMVDR on the CU-Move extended digit task [27,28,37] using our SONIC [23,25] LVCSR system. Sonic incorporates

speaker adaptation and normalization methods such as Maximum Likelihood Linear Regression (MLLR), Vocal Tract Length Normalization (VTLN), and cepstral mean & variance normalization. In addition advanced language-modeling strategies such as concept language models are also incorporated into the toolkit.

The training set includes 60 speakers balanced by age and gender, whereas the test set employs 50 speakers which again are age and gender-balanced. The word error rates (WER) and relative improvements of PMVDR with respect to MFCC are summarized in Table 2-2.

Gender/Sys.	MFCC	PMVDR	Rel. Imp [%]
Female	9.16	5.57	39.2
Male	13.22	8.76	33.7
Overall	**11.12**	**7.11**	**36.1%**

Table 2-2. WERs[%] and relative improvements for CU-Move task

The optimal settings for this task were found to be $M = 24$ and $\alpha = 0.57$ (close to the Bark scale). The 36.1% reduction in error rate using PMVDR features is a strong indicator of the robustness of these features in realistic noisy environments. We tested these features on a number of other tasks including clean, telephone and stressed speech and consistently obtain better results than that for MFCCs. Therefore, we conclude that PMVDR is a better acoustic front-end than MFCC for ASR in car environments.

4.3.5 Integration of Vocal Tract Length Normalization (VTLN)

VTLN is a well-known method of speaker normalization in which a customized linear warping function in the form of $\hat{f} = \beta f$ in frequency domain is used for each speaker [43]. The normalization factor, β, is a number which is generally less than 1.0 for female speakers and more than 1.0 for male speakers to account for different average vocal tract lengths. The normalization factor is determined by an exhaustive search as the one maximizing the total likelihood of a speaker's data using specifically trained models containing only 1 Gaussian for each phoneme cluster for a decision-tree state clustered HMM setting. The VTLN integrated with PMVDR

requires *two consecutive warpings;* one for VTLN and one for incorporation of perceptual considerations.

Gender/Sys.	No VTLN	VTLN	BISN
Female	5.57	4.08	4.25
Male	8.76	7.17	7.10
Overall	**7.11**	**5.57**	**5.62**

Table 2-3. WERs [%] for Speaker normalization performance on CU-Move Corpus

In the PMVDR formulation, we used a first order system to perform perceptual warping. This warping function can also be used for speaker normalization in which the system parameter is adjusted to each speaker [44]. Rather than performing two consecutive warpings, we could simply change the degree of warping, (i.e., α), specifically for every speaker. This will enable us to perform both VTLN and perceptual warping using a *single warp.* The estimation of the VTLN-normalizing α can be done the same way as β. Such an integration of VTLN into the PMVDR framework yields an *acoustic front-end with built-in speaker normalization* (BISN). Table 2-3 summarizes our results with the conventional VTLN and BISN in the PMVDR framework.

The BISN yields comparable results to VTLN with a less complex front-end structure hence is an applicable speaker normalization method in ASR. The total WER reduction compared to the MFCC baseline is around 50% using PMVDR with BISN. The average warping factor for females was α_f=0.55 and for males α_m=0.59. Females require less warping than males due to shorter vocal tract length which conforms to VTLN literature.

Finally, experiments here were conducted on raw speech obtained from one microphone in our array. Using array processing techniques discussed in Sec. 4.1 and integrating the noise information obtained using techniques discusses in Sec. 4.2 will boost performance considerably when used in cascade with the robust acoustic front-end (PMVDR) and built-in speaker normalization (BISN). It is also possible and feasible to apply noise adaptation techniques such as Jacobian adaptation and speaker adaptation techniques such as MLLR to further improve performance[28]. Front-end speech enhancement schemes before acoustic feature extraction was also found to be useful in improving performance [28].

4.4 Proto-type Navigation Dialogue

Finally, we have developed a prototype dialog system for data collection in the car environment [46]. The dialog system is based on the DARPA Galaxy Communicator architecture [47,49] with base system components derived from the CU Communicator system [1,17]. Users interacting with the dialog system can enter their origin and destination address by voice. Currently, 1107 street names for Boulder, Colorado area are modeled. The dialog system automatically retrieves the driving instructions from the internet using an online WWW route direction provider. Once downloaded, the driving directions are queried locally from an SQL database. During interaction, users mark their location on the route by providing spoken odometer readings. Odometer readings are needed since GPS information has not yet been integrated into the prototype dialog system. Given the odometer reading of the vehicle as an estimate of position, route information such as turn descriptions, distances, and summaries can be queried during travel (e.g., "What's my next turn", "How far is it", etc.).

The system uses the University of Colorado SONIC [23,25,48] speech recognizer along with the Phoenix Parser[1] for speech recognition and semantic parsing. The dialog manager is mixed-initiative and event driven [1,17]. For route guidance, the natural language generator formats the driving instructions before presentation to the user by the text-to-speech (TTS) server. For example, the direction, "Park Ave W. becomes 22nd St." is reformatted to, "Park Avenue West becomes Twenty Second Street". Here, knowledge of the task-domain can be used to significantly improve the quality of the output text. The TTS system is based on variable-unit concatenation of synthesis units. While words and phrases are typically concatenated to produce natural sounding speech, the system can back off to smaller units such as phonemes to produce unseen words.

5. DISCUSSION

In this study, we have considered the problem of formulating an in-vehicle speech dialogue system for route navigation and planning. We discussed a flow diagram for our proposed system, CU-Move, and presented results from several sub-tasks including development of our microphone array CSA-BF processing scheme, environmental sniffing, speech enhancement processing,

robust PMVDR features with built-in vocal tract length normalization, and a proto-type dialogue interface via the WWW. We also discussed our speech data corpus development based on Phase I: In-Vehicle Acoustic Noise measurements and Phase II: speech/speaker dialogue collection. Clearly, a number of challenges exist in the development and integration of a natural interactive system in such diverse and changing acoustic conditions. We believe that the processing tasks and results presented reflect useful steps in both the formulation of the CU-Move speech system, as well as contributing to a better scientific understanding of how to formulate dialogue systems in such adverse conditions. Finally, while the prospect of natural hands-free dialog within car environments is a challenging task, we feel that true fundamental advances will only occur if each of the processing phases are capable of sharing knowledge and leveraging their individual contributions to achieve a reliable overall working system.

REFERENCES

[1] W. Ward, B. Pellom, "The CU Communicator System," Proc. IEEE Work. Auto. Speech Recog. & Under., Keystone Colorado, 1999.

[2] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. Signal Processing,* **39**(4):795-805, 1991.

[3] B. Pellom, J.H.L. Hansen, "An Improved Constrained Iterative Speech Enhancement Algorithm for Colored Noise Environments," IEEE Trans. Speech & Audio Proc., **6**(6):573-79, 1998.

[4] P. Lockwood, J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), HMMs and the projection, for robust speech recognition in cars," *Speech Communication,* **11**:215-228, 1992.

[5] S. Riis,O.Viikki, "Low Complexity Speaker Independent Command Word Recognition in Car Environments, IEEE ICASSP-00, **3**:1743-6, Istanbul, Turkey, 2000.

[6] J. Huang, Y. Zhao, S. Levinson, "A DCT-based Fast Enhancement Technique for Robust Speech Recognition in Automobile Usage," EUROSPEECH-99, **5**:1947 –50, Budapest, Hungary, 1999.

[7] R. Bippus, A. Fischer, V. Stahl, "Domain Adaptation for Robust Automatic Speech Recognition in Car Environments," EUROSPEECH-99, **5**:1943-6, Budapest, Hungary, 1999.

[8] A. Fischer, V. Stahl, "Database And Online Adaptation For Improved Speech Recognition In Car Environments," IEEE ICASSP-99, Phoenix, AZ, 1999.

[9] L.S. Huang, C.H. Yang, "A Novel Approach to Robust Speech Endpoint Detection in Car Environments," IEEE ICASSP-00, **3**:1751-4, Istanbul, Turkey, 2000.

[10] E.Ambikairajah, G. Tattersall, A. Davis, "Wavelet Transform-based Speech Enhancement," ICSLP-98, **7**:2811-14, Sydney, Australia, 1998.

[11] P. Gelin, J.-C. Junqua , "Techniques for Robust Speech Recognition in the Car Environment," EUROSPEECH-99, **6**:2483-6, Budapest, Hungary, 1999.
[12] http://www.speechdat.com/SP-CAR/
[13] P. Pollák, J. Vopièka, P. Sovka, "Czech Language Database of Car Speech and Environmental Noise," EUROSPEECH-99, **5**:2263-6, Budapest, Hungary, 1999.
[14] P. Geutner, M. Denecke, U. Meier, M. Westphal, A. Waibel, "Conversational Speech Systems For On-Board Car Navigation and Assistance," ICSLP-98, paper #772, Sydney, Australia, 1998.
[15] M. Westphal, A. Waibel, "Towards Spontaneous Speech Recognition for On-Board Car Navigation and Information Systems," EUROSPEECH-99, **5**: 1955-8, Budapest, Hungary, 1999.
[16] J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," Speech Comm., pp 151-170, Nov. 1996.
[17] B. Pellom, W. Ward, S. Pradhan, "The CU Communicator: an Architecture for Dialogue Systems," ICSLP-2000, Beijing, China, Oct. 2000.
[18] J.F. Kasier, "On a Simple Algorithm to Calculate the 'Energy' of a Signal", IEEE ICASSP-90, pp. 381--384, 1990.
[19] X. Zhang, J.H.L. Hansen, "CSA-BF: Novel Constrained Switched Adaptive Beamforming for Speech Enhancement & Recognition in Real Car Environments", IEEE ICASSP-03. pp. 125-128, Hong Kong, China, April 2003.
[20] P. L. Feintuch, N. J. Bershad, and F. A. Reed, "Time delay Estimation Using the LMS Adaptive Filter-Static Behavior", *IEEE Trans. Acoustics, Speech, Signal Proc.,* ASSP-29(3):571--576, June 1981.
[21] J.H.L. Hansen, et.al., "CU-Move": Analysis & Corpus Develop. for Interactive In-vehicle Speech Systems", Eurospeech-01, pp. 2023-2026, Aalborg, Denmark, 2001.
[22] http://www.nist.gov
[23] Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer", *University of Colorado, Technical Report #TR-CSLR-2001-01,* Boulder, Colorado, March, 2001.
[24] M. Akbacak, J.H.L. Hansen, "Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems," IEEE ICASSP-2003, pp. 113-116, Hong Kong, China, April 2003.
[25] B. Pellom, K. Hacioglu, "Recent Improvements in the CU Sonic ASR System for Noisy Speech," ICASSP-2003, Hong Kong, China, April 2003.
[26] J.H.L. Hansen, "Getting Started with the CU-Move Corpus", Release 2.0A Technical Report, 44pgs., Nov. 17, 2002 [see http://cumove.colorado.edu/].
[27] http://cumove.colorado.edu/
[28] U. Yapanel, X. Zhang, J.H.L. Hansen, "High Performance Digit Recognition in Real Car Environments," ICSLP-2002, vol. 2, pp. 793-796, Denver, CO.
[29] http://speechfind.colorado.edu/
[30] http://speechbot.research.compaq.com/
[31] J.H.L. Hansen, C. Swail, A.J. South, R.K. Moore, H. Steeneken, E.J. Cupples, T. Anderson, C.R.A. Vloeberghs, I. Trancoso, P. Verlinde, "The Impact of Speech Under 'Stress' on Military Speech Technology," NATO RTO-TR-10, AC/323(IST)TP/5 IST/TG-01, March 2000.
[32] M. Akbacak, J.H.L. Hansen, "Environmental Sniffing: Robust Digit Recognition for an In-Vehicle Environment," Eurospeech-03, pp. 2177-2180, Geneva, Switzerland, Sept. 2003.

[33] M. J. Hunt, "Spectral Signal Processing for ASR", Proc ASRU'99, Keystone, Colorado , USA
[34] L. Gu and K. Rose, "Perceptual Harmonic Cepstral Coeffiecients as the Front-End for Speech Recognition", ICSLP-00, Beijing, China, 2000.
[35] M. Jelinek and J.P. Adoul, "Frequency-domain Spectral Envelope Estimation for Low Rate Coding of Speech", IEEE ICASSP-99, Phoenix, Arizona, 1999.
[36] M.N. Murthi and B.D. Rao, "All-pole Modeling of Speech Based on the Minimum Variance Distortionless Response Spectrum", IEEE Trans. Speech & Audio Processing, May 2000.
[37] U.H. Yapanel and J.H.L. Hansen, "A New Perspective on Feature Extraction for Robust In-vehicle Speech Recognition", Eurospeech-03, pp.1281-1284, Geneva, Switzerland, Sept. 2003.
[38] S. Dharanipragada and B.D. Rao, "MVDR-based Feature Extraction for Robust Speech Recognition", IEEE ICASSP-01, Salt Lake City, Utah, 2001.
[39] U.H. Yapanel and S. Dharanipragada, "Perceptual MVDR-based Cepstral Coefficients for Noise Robust Speech Recognition", IEEE ICASSP-03, Hong Kong, China, April 2003.
[40] U.H. Yapanel, S. Dharanipragada, J.H.L. Hansen, "Perceptual MVDR-based Cepstral Coefficients for High-accuracy Speech Recognition", Eurospeech-03, pp. 1425-1428, Geneva, Switzerland, Sept. 2003.
[41] S.L. Marple, Jr, "Digital Spectral Analysis with Applications", Prentice-Hall, Englewood Cliffs, NJ, 1987
[42] K. Tokuda, T. Masuko, T. Kobayashi, and S. Imai, "Mel-generalized Cepstral Analysis-A Unified Approach to Speech Spectral Estimation", ICSLP-94, Yokohama, Japan, 1994.
[43] L.F. Uebel and P.C. Woodland, "An Investigation into Vocal Tract Length Normalization", Eurospeech-99, Budapest, Hungary, 1999.
[44] J. McDonough, W. Byrne, and X. Luo, "Speaker Normalization with All-pass Transforms", ICSLP-98, Sydney, Australia, 1998.
[45] S.E. Bou-Ghazale, J.H.L. Hansen, "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress," IEEE Trans. Speech & Audio Proc., **8**(4): 429-442, July 2000.
[46] B. Pellom, W. Ward, J.H.L. Hansen, K. Hacioglu, J. Zhang, X. Yu, S. Pradhan, "University of Colorado Dialog Systems for Travel and Navigation", in Human Language Technology Conference (HLT), San Diego, California, March, 2001.
[47] URL: Galaxy Communicator Software, http://communicator.sourceforge.net
[48] URL: University of Colorado SONIC LVCSR System http://cslr.colorado.edu/beginweb/speech_recognition/sonic.html
[49] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, V. Zue, "Galaxy-II: A Reference Architecture for Conversational System Development," *Proc. ICSLP,* Sydney Australia, Vol. 3, pp. 931-934, 1998.
[50] X. Zhang, J.H.L. Hansen, "CSA-BF: Novel Constrained Switched Adaptive Beamforming for Speech Enhancement & Recognition in Real Car Environments", IEEE Trans. on Speech & Audio Processing, vol. 11, pp. 733-745, Nov. 2003.