ABSTRACT

In this paper, we present our recent work in the formulation of a new in-vehicle interactive system for route planning and navigation. The novel aspects presented include the formulation of a new microphone array and multi-channel noise suppression front-end, corpus development for speech and acoustic vehicle conditions, environmental classification for changing in-vehicle noise conditions, and a back-end dialog navigation information retrieval sub-system connected to the WWW. 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. Since previous studies in speech recognition have shown significant losses in performance when speakers are under task or emotional stress, it is important to develop conversational systems that minimize operator stress for the driver. In addition to discussing the overall CU-Move system, we report on specific research accomplishments in microphone array/beamforming development, environmental noise characterization, and our prototype dialogue system which is based on the MIT Galaxy-II Hub architecture. We also discuss CU-Move corpus development issues.

1. Introduction

This paper presents our recent work in the formulation of new in-vehicle interactive system for route planning and navigation. This effort is a joint collaboration between CSLR at the Univ. of Colorado Boulder and HRL Laboratories. The system will employ 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 will be an in-vehicle, naturally spoken dialog system to obtain real-time navigation and route planning information using GPS and information retrieval from the WWW. A prototype in-vehicle system is being 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 paper include the formulation of a new microphone array and multi-channel noise suppression front-end, environmental classification for changing in-vehicle noise conditions, and a back-end navigation information retrieval sub-system. We also discuss aspects of corpus development. Most multi-channel data acquisition systems focus merely on standard delay-and-sum beamforming methods. The new noise robust speech processing system uses a five-channel array with a sixth reference microphone placed behind the driver's seat in the automobile. The noise suppression front-end uses beamforming concepts coupled with a frequency dependent partitioning scheme. Specifically, speech data below a frequency cutoff is used directly from the microphone array; data above this cutoff employs a multiplexing beamforming scheme. In essence, a secondary noise reference channel is generated using the array with filter partitions above and below the cutoff. This beamformed reference channel is used with the 6th reference microphone channel, and employed within a constrained iterative speech enhancement processor [2,3] to 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 paper is organized as follows. In Sec. 2, we discuss previous approaches to speech recognition / dialogue in car environments. We also consider research issues and the proposed system flow diagram. Sec. 3 presents work on corpora development. Sec. 4 covers progress in sub-system development and evaluation. Sec. 5 concludes with a summary and discussion of future work.

2. Background & Research Issues

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.) as well as the acoustic environment (road/wind noise from windows, air conditioning, engine noise, exterior traffic, etc.).

† This work was supported by DARPA through SPAWAR under Grant No. N66001-002-8906, and from generous support from HRL Laboratories through CSLR Center Membership.
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; there is significant room for improving hands-free speech collection and the degree of naturalness (i.e., level of scripting) for navigation information exchange.

In developing the CU-Move system, there are a number of research challenges which must be overcome 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 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. Fig. 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. 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 progress in multi-channel noise suppression, automatic environmental characterization, and a prototype navigation dialogue.

3. Data Corpus Development
As part of the CU-Move system formulation, a two phase data collection plan has been initiated. Phase I focuses on collecting acoustic noise and probe speech from a variety of cars and driving conditions. Phase II focuses on an extensive speaker collection across multiple U.S. sites. Fig. 1 shows images of the data collection setup for the multi-channel microphone array and reference microphone. A total of eight vehicles have been selected for acoustic noise analysis. These include the following: compact car(Cmnt), minivan(Mvan), cargo van, sport utility vehicle(SUV), compact and full size trucks, sports car, full size luxury car. A fixed 10 mile route through Boulder, CO was used for
Phase I data collection. The route consisted of city (25 & 45mph) and highway driving (45 & 65mph). The route included stop-and-go traffic, and prescribed locations where driver/passenger windows, turn signals, wiper blades, air conditioning were operated. Each data collection run per car lasted approximately 35-45 minutes. A selection of phonetically balanced TIMIT sentences were read at the same route locations. The following measurements were also recorded: sound pressure level readings, array and reference microphone spacing/location, speaker head location, road and weather conditions. Finally, a video camera was also used to obtain a record of traffic and road conditions during collection. Our plan is to begin Phase II speech/dialogue data collection during fall 2000, which will include (i) phonetically balanced utterances, (ii) task-specific vocabularies, (iii) natural extemporaneous speech, and (iv) conversational interaction with CU-Communicator[1] and CU-Move systems.

4. Sub-System Development & Evaluation

In this section, we discuss our progress in the formulation of a our microphone array/beamformer front-end, environmental classification, speech enhancement processing, and prototype dialogue system.

4.1 Microphone Array & Beamformer

A five-channel microphone array has been constructed using Knowles microphones and a multi-channel data recorder housing built (Fostex) for in-vehicle data collection. Fig. 1 shows the constructed microphone array and data recorder housing. Let us consider the beamformer formulation.

Let \( s(x,t) \) be a signal that is a function of spatial position \( x \) and time \( t \). We can perform a wavenumber-frequency filtering operation (i.e., a 4-D continuous convolution) in both dimensions to obtain the desired response. Thus, the signal received at the \( i \)th sensor can be represented as \( r(x,t) = s(x,t) \). We employ a weighted delay-and-sum beamformer by averaging over the number of sensors:

\[
bf(t) = \frac{1}{N} \sum_{i=1}^{N} w_i x(t - \tau_i)
\]

where \( N \) is the number of sensors, \( w_i \) is the weight of the \( i \)th sensor, and \( \tau_i \) is the delay of the \( i \)th sensor. Sensor weights are based on a trade-off between main beam width and sidelobe power. The coefficients of a Chebyshev polynomial are used since equi-ripple passband criteria can avoid distorting signal components that fall within the main beam. For speech acquisition this is very important, since the mouth, nose, and throat do not behave as a point source when in close proximity to the microphone array. The inter-sensor spacing must be less than half the shortest wavelength component in the signal to avoid passing high frequency components from an undesirable direction.

Construction of the array requires that it contain many sensors to satisfy both the non-aliasing and effective beamwidth conditions simultaneously. However, in the constrained car environment, using many microphones is simply not an option. For our system, an array of five microphones with an inter-sensor spacing of 4.25cm was chosen so as to satisfy the non-aliasing condition, while minimizing the size of the array.

Microphone array sensitivity calibration measurements were performed in the car, with differences between theoretical weighted delay-and-sum beamformer response for frequencies of 1, 3, and 4 kHz and measured responses of less than 5% within an incident angle of 30 degrees. Our experimental results have therefore confirmed algorithm processing with the array to be very close to expected results.

The effect of Chebyshev shading of sensor weights is demonstrated in Fig. 2. The plots show no shading and shading according to a 0.1dB equi-ripple polynomial, respectively. Considerable sidelobe suppression is attainable with this method. Our present work is focusing on integrating the beamformer with a frequency splitting scheme and speech enhancement processing from Sec. 4.3.

4.2 Environmental Classification

One critical processing task for the front-end portion of CU-Move is the environmental classifier. This processor acts as a "sniffer" to determine changing environmental conditions within the car environment, and to direct whether front-end speech enhancement should be engaged, or if the noise condition is more long term (such as windows being opened).
formulated baseline environmental classifier is based on a Gaussian mixture model (GMM) structure. The final implementation will also include objective measure scoring. However, here we restrict the system to use a basic HMM structure. To test the performance, a 4-state HMM was constructed for environmental noise classification. While more than 20 noise conditions have been identified for labeling (some include combinations such as: windows open with turn signal on; external traffic noise with AC on), we performed an initial evaluation using three noise conditions (1. highway acceleration, 2. turn signal on, 3. windows open linc at 65mph) from three vehicles (Cmpt, Mvan, SUV). Sample power spectra from these conditions are shown in Fig. 3. The classifier used a 4-state HMM with 8 MFCCs, energy, and 2 mixtures per state. Average environmental noise classification rates were 84.4%, with the majority of errors made between turn-signal and acceleration conditions. It is important to note that as the number of noise conditions is increased, the likelihood that multiple noise sources becoming active simultaneously will increase.

Fig. 2. Effect of Chebyshev Weights (4kHz signal & 0.1dB ripple)

Fig. 3. Sample acoustic noise power spectra of 3 vehicles. Noise readings across 4kHz bandwidth shown for Windows Open on highway 65mph, and turn signal on at stoplight.

4.3 Front-End Speech Enhancement

While HMM model adaptation will be employed extensively within CU-Move, formulation of the front-end beamforming array has made it attractive to integrate a front-end speech enhancement algorithm. This method offers the advantage of trade-offs in noise suppression across the frequency domain, since beamforming in close proximity is more effective for high frequencies (due to restrictions on microphone count and placement), while speech enhancement can be effective for broadband stationary noise sources at low frequencies. Here, we consider the constrained iterative Auto-LSP speech enhancement method [2,3] with a noise dependent frequency partition (we select the noise partition to be 1.5kHz here). Fig. 4 shows a sample speech spectrogram of the TIMIT sentence “She had your dark suit in greasy wash water all year” spoken within the SUV at 65mph, windows closed. For the original noisy speech, high noise levels are present below 1.8kHz, while good structure is maintained for high frequency fricatives. For the Auto-LSP enhanced result, significant noise reduction is obvious, with the resulting speech sounding much improved with little artifacts. Further improvement will clearly be possible when the beamforming array is integrated within this enhancement method (a factor which we have seen for the case of the “turn signal” noise).

4.4 Proto-type Navigation Dialogue

Finally, we have developed a prototype dialog system for data collection in the car environment. The dialog system is based on the MIT Galaxy-II Hub architecture with base system components derived from the CU Communicator system [1]. Users interacting with the dialog system can enter their origin and destination address by voice. Currently, 1107 street names for Boulder, CO 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 CMU Sphinx-II speech recognizer along with the Phoenix Parser [1] for speech recognition and semantic parsing. The dialog manager is mixed-initiative and event driven [17]. For route guidance, the natural language generator formats the driving instructions before presentation to the user by the text-to-speech 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, beamformer, environmental classifier, speech enhancement processing, and a proto-type dialogue interface via the WWW. We also discussed progress in 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.


