ABSTRACT

In this paper, we present our recent work in the analysis and formulation of a new acoustic speech corpus for developing in-vehicle interactive systems for route planning and navigation. The CU-Move Corpus development is partitioned into two phases: [I] acoustic noise collection and analysis across vehicles, and [II] data collection consisting of +1000 speakers from across the United States. We present results from Phase I acoustic noise data analysis across vehicles to determine guidelines for Phase II large-scale data collection using a single vehicle type. A total of 14 noise conditions are identified for analysis across 6 vehicles. We also present our plan for Phase II collection including speakers, dialect regions, data collection hardware, prompts and dialog domains. Since previous studies in speech recognition have shown significant losses in performance when speakers are under task stress, it is important to develop conversational systems that minimize operator stress for the driver. This will be the first U.S. based corpus of its kind consisting of multi-channel data, intended for use in developing mixed-initiative dialog speech systems; the initial application being route planning and navigation through a wireless information retrieval sub-system connected to the WWW.

1. Introduction

The field of interactive speech systems based on mixed-mode initiative dialog is emerging into new application domains beyond simple command-and-control scenarios and focusing more on fully interactive information access/exchange. The Center for Spoken Language Research has been engaged in the formulation of new in-vehicle interactive system for route planning and navigation[1,2]. The system employs a number of speech processing sub-systems previously developed for the DARPA CU-Communicator[3] (i.e., natural language parser, speech recognition, confidence measurement, text-to-speech synthesis, dialog manager, natural language generation, audio server). The 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 system has been developed for speech corpora collection and system development. This development includes robust front-end processing and recognition model adaptation, as well as a back-end information server to obtain interactive automobile route planning information from WWW.

In this paper, we focus on our two phase CU-Move corpus development effort. Phase I focuses on collection of acoustic data from many vehicles across a wide range of car and road conditions. The motivation is to establish an acoustic corpus to assess the range of vehicle noise conditions for developing novel speech systems. Since such speech recognition development requires data, it is simply not feasible to collect training data in every possible vehicle or road condition. We also note that many countries have laws restricting wireless cell-phone use by drivers. Therefore, it is important to develop reliable low task stress hands-free speech systems that allow natural and relaxed interaction. Clearly, applications such as hands-free wireless communications in vehicles are not new, with numerous studies focused on developing more effective speech recognizers in such environments. However, most studies have focused on isolated-word command and control scenarios for hands-free voice dialing. Studies have shown that even mild levels of user task stress can impact speech production and cause reduction in speech recognition performance[4]. For Phase II, the focus shifts from acoustic noise collection across many vehicles, to an extensive and diverse data collection effort across the United States. While this effort will include navigation commands, digit strings, phonetically balanced sentences, street and road names, a major advantage is an online Wizard-of-Oz dialog interaction that will provide real-time route navigation to the subject in the field.

2. Background & Research Issues

The problem of interactive dialog within car environments offers important speech research challenges. Speech recognition in car environments is fragile, with word-error-rates ranging from 30-65% depending on road and vehicle conditions. These environmental conditions include speaker changes (task stress, emotion, Lombard effect) and acoustic environments (road/wind noise from windows, air conditioning, wiper blades, engine noise, exterior traffic, road surface conditions, weather conditions).

Recent approaches to speech recognition in car environments include combinations of an HMM recognizer with front-end noise suppression[8,9], environmental noise adaptation, and multi-channel concepts. Many early approaches to speech recognition in the car focused on isolated commands[10]. Other studies have shown improvement in computational requirements with front-end signal-subspace enhancement[11]. Another study[12] considered recognizer mismatch between training and testing using clean data and added car noise. Results showed that starting with simulated noisy environment trained models requires about twice as much adaptation material when compared with clean reference models. The work was later extended to consider unsupervised online adaptation[13]. Endpoint detection in car environments[14], and preliminary speech/noise detection with front-end speech enhancement methods as noise suppression for robust speech recognition have also shown promise[8,9,15]. Finally, it is clear that microphone type and placement for in-vehicle speech collection can impact the level of acoustic background noise and ultimately speech recognition performance. has also been considered[14].

† This work was supported by DARPA through SPAWAR under Grant No. N66001-002-8906, and through CSLR Center Membership from HRL, Motorola, and INTEL Laboratories.
Recent work has also been devoted to speech data collection in car environments including the European SpeechDat.Car project[5,6] and others [7]. While most speech data collection focuses on isolated word scenarios such as hands-free voice dialing, there has been some limited work on dialog in cars[17]. The SpeechDat.Car is quite extensive, with data collected in nine languages consisting of navigation words, digits, dates and times, phonetically balanced sentences, and spontaneous sentences. Each speaker session consists of 129 utterances with a total of 600 sessions per database. In their study, the focus is on seven environmental conditions: ranging from stopped car with engine running to driving on highway with radio on; each environment condition is represented by at least 10% of all sessions in the database. A total of 300 speakers are recruited, balanced with respect to age and gender and regional accent. The speaker at times drives the car, or is a passenger depending on regional laws within each country. The recording platform for SpeechDat.Car uses multi-channel system with a 16kHz sample rate on regional laws within each country. The recording platform includes a 4-channel microphone array (Knowles microphone with U.S. coin), (iv) constructed multi-channel DAT recorder (Fostex) with channel dependent level control, (ii) constructed multi-channel digital recorder (Fostex) and individual channel pre-amp controls (Shure), which samples and stores directly to hard-disk at a 44kHz rate. A 5-channel microphone array developed and constructed at CSLR. A reference microphone is positioned behind the driver’s seat. The constructed array and reference microphone use Knowles microphones (an example is shown next to U.S. coin in Fig. 1). An additional far-talking microphone and cell-phone channel are also used in the data recording setup. A flat-panel display with laptop is used to prompt speakers for the individual speech areas for Phase II.

In Phase I data collection, six vehicles were selected for acoustic noise analysis. A prescribed 17 mile route was selected in Boulder, CO. The route consists of city (25 & 45mph) and highway driving (45 & 65mph); this includes stop-and-go traffic, prescribed locations where driver/passenger windows were opened at different amounts, and operation of turn signals, wiper blades, and air conditioning. Each data collection run per car lasted approximately 40 minutes. Six cars were used for analysis: compact car(Cavalier), minivan(Ventura), cargo van (Express), sport utility vehicle(Blazer), compact (S10) and full size (Silverado) trucks. A selection of phonetically balanced TIMIT sentences were read at the same route locations. All data was transcribed at three levels: (i) text level, (ii) car dependent noise (i.e., turn signals, wiper blades, acceleration, etc.), and (iii) non-dependent car noise (i.e., passing traffic, exterior road or weather conditions, etc.). We determined 14 noise conditions to be labeled in the data. Fig. 2 shows sample speech spectrograms of turn signal noise from the six vehicles with windows closed. In our analysis, we considered average signal power between 0-1.5kHz versus 1.5-4kHz regions. Results show that for noise such as turn signal, there is some variability across vehicles. The average impulse rate for the turn signal varied from 2.75 – 3.00Hz, with sharp differences on how distinct the impulse points are from background noise (as seen in Fig. 2). In order to determine noise levels versus speech, an analysis was performed to determine average log power differences between (speech + noise) and noise. Fig. 3 shows frequency dependent differences for one sentence spoken while windows were open 1-2inches traveling at 65mph. It is clear that the level of background noise is very close to the speech level in the frequency band 0-1kHz (avg. of 2.55-2.65dB difference). The background noise level drops measurably in the frequency range between 2-4kHz, with an average speech level between 13.8-17.2dB above the background noise floor. While there is some variability across the six vehicles here, it is encouraging that noise levels are generally within one standard deviation of the mean for all cars. There was more consistency across vehicles for windows closed traveling 65mph, and more variability with windows open half-way. The frequency bin

3. Collection Platform & Phase I Analysis

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. cities. Fig. 1 shows images of the data collection setup. The data collection platform consists of a constructed data recorder housing with multi-channel digital recorder (Fostex) and individual channel...
with the largest differences between speech and noise was at 1kHz, with the mean speech level over noise moving from 10.7dB – 2.65dB – 5.05dB as the windows moved from closed, to 1-2inches open, to half-way open respectively. This suggests that it would be possible to train one HMM model across vehicles for some noise types if the variability is small enough. Finally, while results from Fig. 2 and 3 show variability across vehicles, we use Fig. 4 to show changes in noise conditions for a single vehicle. Here, engine idle with windows closed represent the lowest noise level condition, and windows open 1-2inches traveling 65mph is the noisiest. The majority of the 14 noise sources are concentrated within a 20dB (low freq.) by 10dB (high freq.) grid, the center of which does appear to be vehicle dependent. The noise analysis performed in Phase I allows us to categorize the various noise types and levels across cars, and to predict how appropriate speech data collected in one vehicle might be modified in the training phase for speech recognition in other vehicles. Noise conditions that varied significantly across vehicle types include: 65mph & 45mph (windows open 2in), deceleration-windows open 2in, turn signal-windows closed. Noise conditions that were more tightly grouped across vehicle types include: idle, wipers on-windows closed, 65mph-windows closed. Other noise conditions showed groupings (i.e., trucks and vans close together, cars and SUV close together).

**Fig. 2.** Spectrograms of acoustic noise from vehicles: Turn signal with windows closed at 65 mph: from top to bottom – SUV, Compact car, Extended pickup truck, commercial van, passenger van, compact pickup truck.

**Fig. 3.** Difference between speech+noise and noise levels for 6 vehicles with windows open 2 in. traveling at 65mph.

**Fig. 4.** Average dB noise levels between low (0-1.5kHz) and high (1.5-4kHz) frequency for 14 noise conditions in an SUV vehicle. Noise conditions are: traveling 45mph (windows open 2in, closed, open ½ way), traveling 65mph (windows open 2in, closed), acceleration (windows closed, open ½ way), AC on (windows closed), deceleration (windows closed), turn signal on (65mph – windows closed, windows open 2in), windshield wipers on (windows closed), engine idle (windows closed), engine idle (windows opened), engine idle (windows closed), engine idle (windows open 2in). 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]. 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. The initial 6 month plan (beginning April 1), is to collect speech across 7 U.S. cities that reflect regional dialects. These cities were selected to be mid-size cities (75-150k people), in order to increase the prospects of obtaining subjects who are native to that region. In each city, we will collect between 100-150 speakers, with a minimum of 100 being native to that region. Additional speakers will be collected in each city that might reflect other speaker groups that may not be associated with a particular region. The speaker population will be balanced using figures available from the U.S. Census [http://quickfacts.census.gov/qfd/]. Balance across

4. Data Corpus Development

Next, we shift our discussion to our Phase II speech collection plan. In developing the CU-Move system[1], 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]. 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. The initial 6 month plan (beginning April 1), is to collect speech across 7 U.S. cities that reflect regional dialects. These cities were selected to be mid-size cities (75-150k people), in order to increase the prospects of obtaining subjects who are native to that region. In each city, we will collect between 100-150 speakers, with a minimum of 100 being native to that region. Additional speakers will be collected in each city that might reflect other speaker groups that may not be associated with a particular region. The speaker population will be balanced using figures available from the U.S. Census [http://quickfacts.census.gov/qfd/]. Balance across

* At the present time, seven U.S. cities will be used during Phase II speech collection. Additional cities will be added when further resources become available.
gender and age brackets will also be maintained. The driver will perform 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.) will be included. The format of the data collection will consist of five domains that consist of 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 a right or left turn?” “How do I get to the airport?” “I’m lost. Help me.”

Digit Sequences: each speaker will produce 16 digit strings from a randomized 75 digital string set. Examples include: telephone numbers (733-2034, 425-952-5406), say a familiar phone number), random credit card numbers (1234-5621-1253-5981), and individual numbers (0,0,1,886, *551).

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

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

Dialog Wizard — of — Oz Collection: this section represents one of the most important phases of the CU-Move collection, since it will involve exercising an actual route navigation system. Each speaker will use the in-vehicle cell-phone connection to call our working CU-Move navigation system at CSLR. Since we are collecting speech for speech recognition training and model adaptation development, the speaker will make their request, but a WOZ operator will listen for the requested address and type in the request. They will select at least 3 destinations and call the system from the vehicle, get navigation information and ask general questions about the planned route. Examples include: “driving directions from police station to a gas station, directions from a hotel to a restaurant, etc.

For the WOZ collection, we developed a prototype dialog system for navigation 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 [3]. Users interacting with the dialog system can enter their origin and destination address by voice. Currently, 4100 street names for Boulder, CO area are modeled (note: one reason for using a WOZ in Phase II is to overcome the issue of attempting to anticipate every possible street name in the present language model). 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. The dialog portion will have speech recorded in the vehicle as well as at the CU-Move server at CSLR.

Finally, once data is collected, all sections will be transcribed, with careful consideration given to inter-labeler variability analysis. Transcribers will be dedicated to collection domains, so that multi-channel speech and transcripts can be made available for in-vehicle speech and dialog research.

5. Discussion

In this paper, we have discussed analysis and corpus development for the CU-Move in-vehicle speech dialogue system for route navigation and planning. Corpus development has been partitioned into two phases: I: in-vehicle acoustic noise collection and analysis across vehicles, and II: a national data collection consisting of +1000 speakers from across the United States. Phase I has been completed, and results from noise analysis across multiple vehicle types and 14 driving/car conditions were presented. We saw that some noise conditions are more consistent across vehicles, while others are more vehicle dependent. Also, these 14 noise conditions have been included in a definition format for the LDC publicly available transcription tool. Noise conditions include car-dependent and car-independent noise classes. We also presented our Phase II collection plan (which will be completed by Sept. 2001) that includes speakers, dialect regions, data collection hardware, data prompt and dialog domains. Clearly, a number of challenges exist in developing effective natural interactive speech systems in such diverse acoustic conditions. Clearly a main goal is to minimize user task stress levels if the driver is accessing information such as route navigation. Finally, we point out that other information access scenarios are possible, and the Center for Spoken Language Research is presently exploring such applications as entertainment (voice browsing for audio/MP3 music, sporting events/scores), weather, and traffic conditions.

REFERENCES