THE AUSTRALIAN ENGLISH SPEECH CORPUS FOR IN-CAR SPEECH PROCESSING

Tristan Kleinschmidt, Michael Mason, Eddie Wong, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology
Brisbane, QLD, Australia

{t.kleinschmidt, m.mason, ee.wong, s.sridharan}@qut.edu.au

ABSTRACT

The Australian In-Car Speech Corpus is a multi-channel recording of a series of prompts from an in-car navigation task collected over a range of speakers in a variety of driving conditions. Its purpose is to provide a significant resource of speech data appropriate for investigating speech processing needs in the adverse environment of a car. Utterances spoken by 50 speakers were collected in seven different driving conditions, providing the foundation for investigation into noisy, speaker-independent speech processing. Speech recognition experiments are performed to validate the data, to provide baseline results for in-car speech recognition research, and to show that this data can improve speech recognition performance under adverse in-car conditions for Australian English when adapting from American English acoustic models.

Index Terms— Multimedia databases, speech recognition, road vehicles.

1. INTRODUCTION

Consumer devices such as navigation systems and media players are becoming more common in automotive environments. In order to maintain safety on roads, there is an increasing need to provide intuitive and safe user interfaces to these devices. Considerable improvements in technology have led to in-car speech recognition systems being well positioned to fulfill this need.

The key challenge of deploying speech recognition in real-world environments is the requirement for high accuracy in the presence of high levels of noise. Since most speech recognition systems are trained for use in controlled environments, they fail to produce satisfactory performance under more adverse conditions such as in automotive environments.

One of the major limitations in making speech recognition systems more robust is the requirement to collect sufficient amounts of data on which to train models [1] and perform meaningful evaluations. Both tasks require hundreds of hours of work in collecting data and transcribing it. As a result, training acoustic models for an intended operating environment is often abandoned, and techniques such as model adaptation and speech enhancement are introduced to improve overall system performance on a smaller set of test data.

The second limitation is the availability of only single-channel speech recordings. State-of-the-art speech enhancement techniques (e.g. beamforming or adaptive noise cancelation) use multiple microphones, therefore the lack of multi-channel recordings makes development and comparison to single-channel enhancement algorithms difficult.

Parts of the work presented here were funded through the Australian Cooperative Research Centre for Advanced Automotive Technology (AutoCRC). To acquire a copy of this database, contact the last named author.

In order to overcome these limitations, a number of large in-car speech databases have been collected [2, 3, 4]. These collections contain recorded speech from a large number of speakers under a wide range of noise conditions. To date no such collections have been performed under Australian driving conditions using Australian accented speakers.

The rest of this paper is organised as follows. Section 2 outlines the collection of the multi-channel Australian In-Car Speech database. Section 3 explains the development of the speech recognition evaluation protocol for the database. In order to validate the data set, Section 4 outlines a baseline speech recognition system and reports on the corresponding recognition performance of the evaluation protocol including model adaptation results.

2. AUSTRALIAN IN-CAR SPEECH DATABASE

2.1. Recording Setup

2.1.1. Equipment

A 2008 VE Commodore was outfitted as part of the Australian Cooperative Research Centre for Advanced Automotive Technology (AutoCRC) for use as the data collection vehicle. The key components of the data collection system were an in-car PC running Microsoft Windows and the LabView software used to record the data. A National Instruments CompactDAQ USB-based data acquisition chassis with two 4-channel analogue acquisition modules (NI WLS-9234) acted as the analogue-to-digital converter (ADC) and multi-channel simultaneous sampling device. A series of custom-designed biasing boxes were used to provide virtual power and impedance matching for eight high-quality Sennheiser MKE 2-5 omnidirectional electret microphones. Collecting high-quality recordings enables the recordings to be degraded to simulate cheaper, lower-quality microphones used in production vehicles.

A custom LabView virtual instrument was used to manage prompts and control the recordings. The PC (with touch screen interface for ease of input) was operated by a research assistant in the front passenger seat whilst the speaker was driving the car.

2.1.2. Microphone Location

The eight microphones were fitted to the central roof console pointing downwards as shown in Fig. 1. This location is an industry-favoured position due to the ease of integration with existing electronics whilst still providing good signal-to-noise ratios [2]. The microphones were spaced symmetrically around the midline of the vehicle with 2 cm between each adjacent microphone. The average location of the driver's mouth was estimated (with reference to the microphone closest to the driver) to be 35 cm to the right, 25 cm below, and 17.5 cm behind this reference microphone.

Table 1. *Extended Backus-Nauer form grammar used in the collection of the Australian In-Car Speech Corpus.*

$Numbers	= [NUMBER] (**$Single_Digit** \| **$Two_Digits** | **$Three_Digits** \| **$Four_Digits**);
$Street	= [**$Street_Prefix_List**] **$Street_Name_List** **$Street_Type_List**;
$In_Suburb	= [(AT | IN)] **$Suburb_List**;
$Corner	= (CORNER | JUNCTION | INTERSECTION) OF **$Street** AND **$Street** [**$In_Suburb**] ;
$Address	= ([**$Numbers**] **$Street** **$Suburb_List**) \| (**$Suburb_List** **$Street** [**$Numbers**]) | **$Corner**;
$Addr_Cmd	= ENTER (ADDRESS | DESTINATION) **$Address**;
$Other_Cmd	= RECALL DESTINATION **$Single_Digit** | (START | STOP) NAVIGATION | RETURN | BACK | MAIN MENU;
$Cmd	= **$Addr_Cmd** \| **$Other_Cmd**;

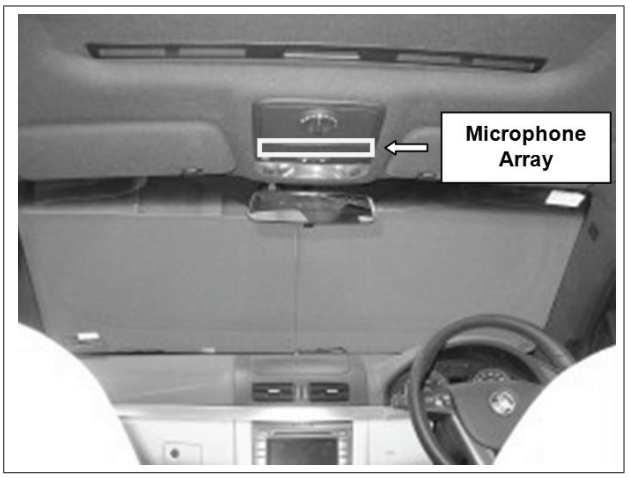

Fig. 1. *Location of 8-microphone array used in Australian In-Car Speech Corpus collection.*

2.1.3. Recording Format

The CompactDAQ ADC modules and LabView virtual instrument recorded the audio data at the natural sampling frequency of 50 kHz and stored the 24-bit sample resolution of the eight channels as 32-bit values. Output from the virtual instrument was a single 8-channel interleaved 23-bit per sample waveform file stored in the WAVE/RIFF file format.

Following acquisition, the primary files were separated into individual channels and re-sampled to 16-bits per sample and a 16 kHz sampling rate. Channels were labelled 0 through 7 sequentially representing the microphone closest to the driver (mic0) to the one closest to the front seat passenger (mic7). Resulting audio files are stored in WAVE/RIFF file format while the accompanying label files containing utterance level transcriptions are stored as text files.

2.2. Speaker Description

A total of 50 speakers are present in the database with 20 female speakers and 30 male speakers represented. Female speakers were aged between 21 and 53 years, male speakers between 20 and 67 years old. Only native English speakers were collected. For accent purposes, it was a requirement that English be the *first* language of the participant, even if English was acquired at such a young age as to be commonly considered "native" to that speaker. Also, it was a requirement that each participant had lived in Australia for at least five years to allow for some Australian accent naturalisation.

Table 2. *Seven in-car noise conditions in the Australian In-Car Speech database.*

Condition	Description
C0	Car idle, sealed cabin, no HVAC
C1	Medium speed (50-60 km/h), sealed cabin, no HVAC
C2	Medium speed (50-60 km/h), sealed cabin, HVAC on full fan
C3	Medium speed (50-60 km/h), driver window open, no HVAC
C4	High speed (90-100 km/h), sealed cabin, no HVAC
C5	High speed (90-100 km/h), sealed cabin, HVAC on full fan
C6	Car idle, sealed cabin, HVAC on full fan

2.3. Task Description

A command-and-control extended Backus-Nauer form (EBNF) style grammar (shown in Table 1) was formulated to generate a large number of consistent utterances for drivers to say in a variety of driving conditions based on a mock navigation task. The grammar was generated using 20 suburbs, 1931 street names, 16 prefixes and 37 street types from the cities of Brisbane and Melbourne. The task-oriented grammar provides the potential to investigate language processing techniques which may aid medium-vocabulary command-and-control applications.

2.3.1. Recording Conditions

Seven different driving conditions were used as general audio scenes for utterance recording. These conditions were chosen to capture variety in general noise types and levels present in the cabin of a vehicle whilst also representing likely driving scenarios in Australia. Table 2 shows a full list of the recording conditions. The acronym HVAC stands for the Heating, Ventilating, and Air Conditioning system. The main vents are located on the centre console directly beneath the microphone array which can be seen in Fig. 1.

2.3.2. Prompt Generation

Each speaker was recorded saying a series of utterances in a number of driving conditions. For each driving condition, the speaker recorded six utterances. The utterances were classified as Common, Repeated and Unique utterances. The Common utterances are a set of utterances which each participant recorded identically. A set of common utterances is associated with each specific driving condition (C0000-C0006). In each condition participants recorded two

Table 3. Speaker groupings used in the Australian In-Car Speech Database evaluation protocol.

Group	Speakers	# Utterances
I	P04, P05, P11, P14, P16 P17, P21, P26, P35, P42	714
II	P08, P09, P12, P15, P22 P27, P30, P34, P47, P49	840
III	P02, P07, P18, P23, P38 P39, P43, P46, P54, P55	790
IV	P10, P19, P24, P25, P31 P32, P36, P45, P52, P53	720
V	P03, P06, P13, P20, P28 P29, P33, P37, P41, P51	749

Repeated utterances. Repeated utterances occur more than once in the entire database, though never twice by the same speaker. Multiple speakers may record the same Repeated utterance in the same driving condition. Each speaker also recorded three Unique utterances per condition. These utterances occur only once across the entire database.

In total, each speaker nominally recorded 42 utterances. The utterances consisted of two different types of information – an address-style utterance, or a chain of commands which would be used in a navigation system. The command utterances contain a chain of 6 commands, never in the same order for Unique utterances. While these command chains were recorded as one utterance, they have been separated into individual files in the final database.

During validation a small number of exceptions caused utterances to be unusable and they have been removed from the database. The final database contains 3787 utterances. A full list of the recorded utterances is included in the database documentation.

3. EVALUATION PROTOCOL

The database outlined in Section 2 is suitable for use in a number of speech processing fields such as speech enhancement and speech recognition. The multiple channel recording process ensures investigations into current beamforming techniques are possible. For single-channel experiments (as per the speech recognition baseline results presented in Section 4) microphone 0 should be chosen as it is closest to the driver and generates the highest recognition accuracies based on our preliminary experiments. Multi-channel techniques can utilise whichever channels are required for the individual technique.

The proposed speech recognition evaluations on the collected data break the 50 speakers down into a series of groups. This allows the data to be used for model adaptation, development and evaluation testing through the use of k-fold leave-one-out testing. Results can be averaged across a desired number of the folds to provide a more indicative speaker-independent word recognition rate. In doing this, the effect of very good (or very bad) performance of individual speakers is reduced.

The 50 speakers have been broken up randomly into five groups of 10 speakers as shown in Table 3. To ensure some level of consistency amongst the groups, an effort was made to ensure a balance of male and female speakers in each group – as a result there are 6 male and 4 female speakers in each grouping.

To facilitate model adaptation, development and evaluation testing, three groups (approximately 60% of the data including all noise conditions) are made available for adaptation, one group for development testing, with the fifth group being used for evaluation testing purposes. Five combinations of this segregation are shown in Table 4. It is intended these groupings be used in the order stated in the table; individual experiments can dictate the number of folds required.

Table 4. Australian In-Car Speech Database protocol groups for k-fold leave-one-out speech recognition experiments.

Fold	Adaptation	Evaluation	Testing
1	I, II, III	IV	V
2	III, IV, V	I	II
3	I, II, V	III	IV
4	II, III, IV	V	I
5	I, IV, V	II	III

4. SPEECH RECOGNITION EXPERIMENTS

In order to provide a common reference to facilitate simple results comparisons and also validate the collected data, two speech recognition experiments have been performed. In this section we define the baseline recogniser used for these experiments, as well as report on results of model adaptation with this data.

4.1. Baseline Recogniser

Context-dependent 3-state triphone hidden Markov models (HMM) were trained using the American English Wall Street Journal 1 corpus to enable speaker-independent speech recognition. The acoustic models were trained using 39-dimensional Mel-Frequency Cepstral Coefficient (MFCC) vectors – 13 MFCC (including C_0) plus delta and acceleration coefficients. Each HMM state was represented using a 16-component Gaussian Mixture Model. Utterance decoding was performed using the Hidden Markov Model Toolkit (HTK) [5]. The grammar used to generate the utterances for the data collection (Table 1) was used as the task grammar for these experiments.

All speech recognition results quoted in this paper are word accuracies (in %) and are calculated as:

$$PercentAccuracy = \frac{N - D - S - I}{N} * 100\% \qquad (1)$$

where N represents the total number of words in the experiment, D the number of deletions, S the number of substitutions and I the number of insertions [5].

4.2. Experimental Results

4.2.1. Baseline Results

Baseline results were generated using the original clean acoustic models trained as per Section 4.1. The results are shown in the top half of Table 5. Results are collated by noise condition and experimental fold, with the average results shown being the combined accuracy over all experimental folds.

Analysing the results in Table 5, a number of observations related to the in-car noise conditions can be made. Comparing the results for all car speeds with either the HVAC on (C2, C5, C6) or off (C0, C1, C4), it can be seen that an increase in vehicle speed causes degradation in the recognition accuracy. The decrease in performance is particularly noticeable in the case where the air-conditioning system is off, where accuracies are 84.89%, 70.22% and 54.12% for idle, 50-60 km/h and 90-100 km/h respectively. This is due to increases in road and wind friction as vehicle speed increases.

Table 5. Baseline and MAP adaptation word accuracy (%) results for the Australian In-Car Speech Database.

	Baseline						
Fold	C0	C1	C2	C3	C4	C5	C6
1	85.06	67.73	31.55	48.25	57.11	25.89	43.75
2	84.65	70.49	36.57	56.05	54.91	26.72	36.83
3	88.40	75.57	31.81	50.57	57.64	31.15	44.62
4	82.40	67.61	35.26	58.30	51.03	38.06	45.23
5	84.05	69.70	38.41	52.59	49.90	35.88	37.60
Aver. 1-5	84.89	70.22	34.74	53.09	54.12	31.61	41.64
	MAP Adaptation						
Fold	C0	C1	C2	C3	C4	C5	C6
1	90.87	83.86	71.24	83.43	85.57	72.98	78.32
2	90.51	87.30	75.35	82.56	84.34	68.13	74.07
3	95.15	90.19	72.37	80.38	87.17	64.25	72.80
4	91.72	86.44	70.94	81.98	78.60	72.57	73.24
5	92.23	85.88	73.98	80.68	80.24	72.29	68.90
Aver. 1-5	92.08	86.76	72.81	81.81	83.19	70.01	73.50

Having the air-conditioning system on appears to have the greatest effect on the recognition accuracy, rather than simply increasing vehicle speed. In the idle case, the performance difference is approximately 43%. Having the windows down (C3) doesn't degrade the performance anywhere near as drastically as the air-conditioning. This phenomenon can be attributed to the location of the air-conditioning vents which are directly beneath the microphone array, and therefore fan noise is recorded by the microphones at considerably higher amplitudes than noise coming from the driver's side window.

4.2.2. Adaptation Results

To test the effectiveness of the Australian In-Car Speech Corpus for adapting clean speech acoustic models, maximum *a posteriori* adaptation (MAP) [1] was chosen. The pre-trained triphone models described in Section 4.1 were assumed to provide a good initial estimate of the parameter distribution required by MAP adaptation. Both mean and variance adaptation has been performed using the recordings from microphone 0, with a value of τ of 16 (i.e. the prior model has 16 times more influence than the adaptation data). This value was chosen since the adaptation set has considerably fewer speakers than the training data, and therefore it is required to ensure the models remain speaker independent. Speech recognition results using the adapted models are shown in the bottom half of Table 5.

The adaptation results show uniform improvements in word accuracy over the baseline results presented in the top half of Table 5 for all speaker groups. The significant increase in performance (on average 29%) could be attributed to two factors – adaptation to the in-car noise conditions and adaptation to the Australian accent. Further, since the prompt generation ensures command-based utterances are said on a regular basis, the model coverage of the adaptation task is very high for these common words used in the evaluation. The effect of each of these factors are the focus of future research into adaptation for in-car speech dialogue systems.

5. CONCLUSION

A new in-car speech corpus collected in Australian driving conditions with Australian accented speakers has been described. Seven recording conditions were chosen to reflect Australian driving scenarios, and the task grammar was chosen to mimic navigation system commands and addresses. This database is suitable for in-car speech recognition evaluations as well as single- and multi-channel speech enhancement algorithms.

For speech recognition evaluations, a k-fold leave-one-out protocol enabling model adaptation, evaluation and testing has been proposed. Baseline speech recognition experiments have validated the data as behaving as expected under in-car noise conditions. MAP adaptation using the proposed framework on speech models trained with American English shows consistent accuracy improvement for all noise conditions over baseline results, demonstrating the feasibility of creating an Australian English speech recogniser based on existing resources.

To acquire a copy of this database, contact the last named author.

6. REFERENCES

[1] C. Lee and J. Gauvain, "Bayesian adaptive learning and MAP estimation of HMM," in *Automatic speech and speaker recognition : Advanced topics*, pp. 83–107. Kluwer Academic Publishers, Boston, Massachusetts, USA, 1996.

[2] M. Sala, H. Wengelnik, H. van den Heuvel, A. Moreno, E. Le Chevalier, E. Deregibus, and G. Richard, "SpeechDat-Car: speech databases for voice driven teleservices and control of in-car applications," in *Proc. EAEC*, Barcelona, 1999, pp. 90–98.

[3] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. INTERSPEECH*, Jeju Island, Korea, 2004, pp. 2489–2492.

[4] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J.H.L. Hansen, "UTDrive: Driver behavior and speech interactive systems for in-vehicle environments," in *IEEE Intelligent Vehicles Symposium*, Istanbul, Turkey, 2007, pp. 566–569.

[5] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 3.4 edition, December 2006.