A Continuous Speech Recognition Evaluation Protocol for the AVICAR Database

Tristan Kleinschmidt, David Dean, Sridha Sridharan, Michael Mason
Speech and Audio Research Laboratory
Queensland University of Technology,
GPO Box 2434, Brisbane, Australia, 4001
{t.kleinschmidt, d.dean, s.sridharan, m.mason}@qut.edu.au

***Abstract*— The use of speech recognition in automotive environments has received increased attention in recent times. Unfortunately, evaluations of algorithms designed to improve recognition performance in this environment have been performed on differing data collections, making results difficult to compare. In recent years, the University of Illinois released a large in-car audio and visual data collection known as AVICAR ("audio-visual speech in a car") [1]. The AVICAR database is freely available, but to date no uniform evaluation protocol on which to perform experiments has been reported. This paper introduces a speaker-independent, continuous speech recognition evaluation protocol for the audio data of the AVICAR database. It is designed to allow for model adaptation, evaluation and testing using native English speakers. Baseline recognition results obtained using this protocol are also presented.**

I. Introduction

The key challenge of deploying speech recognition in real-world environments is the stringent performance 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 ability to collect sufficient amounts of data on which to train models and perform meaningful evaluations. The former task often requires hundreds of hours of work in collecting data and transcribing it. As a result, training acoustic models for the intended operating environment is often abandoned, and techniques such as model adaptation and speech enhancement are introduced to improve overall system performance.

Whilst considerable in-car speech recognition research has been conducted, the data used for experimentation is often collected solely for the individual evaluation. This typically ensures only limited amounts of data are used in the evaluation. It also means that making performance comparisons between proposed techniques is almost impossible unless all techniques in question are evaluated on each data set – an unnecessary and time-consuming process.

To alleviate this issue, the Aurora experimental framework [2] was introduced. This framework uses a noisy version of a large database designed for speaker-independent isolated digit recognition [3]. Noisy speech was created by artificially adding various noise sources to clean speech at levels which satisfied a range of signal-to-noise ratios (SNR). A baseline recognition system was also documented to provide a common platform for straightforward comparisons of results.

Although this database has been used extensively to report and compare experimental results, there are two very important limitations imposed by this particular framework. Firstly, in scenarios where noise signals are synthetically added to clean speech data (as is the case with Aurora), no alteration is made to the speech waveform. Whilst this ensures various SNRs can be achieved, it fails to reflect changes in speech production which occur with increased ambient noise levels. This effect is known as the Lombard effect and has been shown to be an important factor in performance of noisy speech recognition systems [4]. Whilst the presence of the Lombard effect in modern-day vehicles has yet to be properly confirmed or rejected, the only way to answer this question is to use speech collected in a vehicle whilst it is being driven.

The second limitation is the availability of only single-channel recordings. State-of-the-art speech enhancement techniques (e.g. beamforming or adaptive noise cancellation) use multiple microphones – therefore the Aurora framework is unable to be used for evaluation of these techniques. This makes comparison of algorithms even more difficult, particularly when trying to show performance gains through use of multiple microphones.

In order to overcome these limitations, a number of large in-car speech databases have been collected [5], [6]. These collections contain recorded speech from a large number of speakers under an extensive range of noise conditions. Unfortunately, these datasets have been used in few studies because they are either very expensive to acquire or not publicly available.

With the release of the AVICAR ("audio-visual speech in a car") database from the University of Illinois [1], there is potential to create a uniform evaluation protocol for in-car speech recognition using real in-car speech data. This database is freely available so there is potential for widespread use to study methods which further improve in-car speech recognition. Another advantage of the AVICAR database is

Parts of the work presented here were funded through the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

TABLE I

AVICAR DATABASE IN-CAR NOISE CONDITIONS.

Noise	Description
IDL	Engine running, car stopped, windows up
35U	Car travelling at 35mph, windows up
35D	Car travelling at 35mph, windows down
55U	Car travelling at 55mph, windows up
55D	Car travelling at 55mph, windows down

the ability to perform multi-channel experiments, making it an attractive corpus for evaluating state-of-the-art speech enhancement techniques. This paper proposes an evaluation framework for the audio portion of the AVICAR database which enables single- and multi-channel, speaker-independent, continuous speech recognition (CSR) experiments.

The rest of this paper is organised as follows. Section II presents important aspects of the AVICAR database. Section III explains the development of the evaluation protocol. Section IV outlines the baseline recogniser and the corresponding recognition performance of this framework including model adaptation results.

II. AVICAR DATABASE

The AVICAR database is a large, publicly available in-car speech corpus containing multi-channel audio and video recordings. It was recorded by researchers at the University of Illinois. The collection was designed to enable low-SNR speech recognition through combining multi-channel audio and visual speech recognition. Detailed information about the recording setup can be found in [1].

The released portion of the AVICAR database contains less data than documented in [1]. It includes audio for 87 speakers and video for 86 subjects. Around 60% of the speakers are native American English speakers, with the remainder of the speakers being native to Latin America, Europe, East or South Asia. All the recorded speech, however, is in English.

Four distinct tasks exist in this data collection – isolated digits, isolated letters, phone numbers and TIMIT sentences. These four different tasks can be used for a range of speech recognition tests. The isolated digits task closely resembles command-and-control applications, whilst the isolated letters task mimics spelling which may be required in navigation systems. The other two tasks constitute continuous speech recognition tasks – phone numbers represent small-vocabulary systems, whilst sentences match medium-vocabulary systems. Further information about the utterance scripts used in the collection can be found in [1].

Each recording session contains speech under five noise conditions. This enables analysis of the affect on recognition performance of different types of noise from common driving scenarios. The noise conditions are detailed in Table I.

A standard framework for performing speech recognition tests on isolated digits and letters is provided with the release of the database. The scripts utilise pre-trained models (which come with the database) and the Hidden Markov Model Toolkit (HTK) [7] to generate test results. Recognition frameworks are only available for the two isolated word tasks, therefore an evaluation protocol for the continuous speech tasks (i.e. phone numbers and TIMIT sentences) is required. A protocol meeting the requirements for speaker-independent, continuous speech recognition has been devised for the AVICAR database and is outlined in Section III.

III. EVALUATION PROTOCOL

A. Protocol Design

Having access to both microphones and cameras in production vehicles is still a vision for the future. Currently, most interest is in continuous speech recognition in the audio space, particularly with multiple microphones. As such, only the audio portion of the phone number and sentence tasks have been considered in this protocol. Future work should be able to extend this protocol to include the video data or the isolated word tasks.

As stated in Section II, the two tasks used in this protocol provide distinct test scenarios for the in-car environment. Phone numbers enable testing of small-vocabulary CSR, and sentences medium-vocabulary CSR. The sentences task is related to command-and-control and navigation applications which are both important tasks for in-car speech recognition.

To create a protocol which enables uniform adaptation and testing of a range of single- and multi-microphone speech enhancement techniques, a number of restrictions needed to be put on the data.

1) An individual utterance must have all 7 microphones in the array available with valid audio (i.e. the eighth microphone connected to the camcorder is excluded).
2) A speaker must be a native English speaker (this includes British and American English) *and* have at least one utterance satisfying criteria 1.

To ensure the first criteria is met, a list of discarded files included with the database was consulted to remove unusable utterances. The discard list details recordings which have some (but not all) microphone audio missing or corrupted due to hardware problems. Some utterances have no audio data at all, but are not included in the discard list – therefore the list was examined simultaneously with data that actually exists in the database release.

The second criteria is met by analysing metadata. Grouping available utterances by speaker confirmed that 55 speakers were suitable for use in the evaluation protocol.

In order to be consistent with requirements for model adaptation, as well as system tuning and testing, it was decided to split the 55 speakers randomly into five groups of 11 speakers. To ensure some level of consistency amongst the groups an effort was made to evenly balance male and female speakers as well as an even distribution of the utterance scripts. A sixth group (denoted Group VI) was created solely for non-native English speakers using the same rules described above. The resulting six speaker groups are listed in Table II.

A primary goal of speech enhancement research for in-car speech recognition is to analyse speaker-independent performance across different noise levels and conditions. With

TABLE II

AVICAR DATABASE PROTOCOL SPEAKER GROUPS.

Group	Speakers
I	AM4, BM4, CF5, DF1, EF4, EM1, FF2, GM4, HF2, HM3, IF1
II	AM3, BF5, BM1, CM1, DM2, EM4, FF5, GF2, HF3, IM5, JM2
III	AM2, BM3, CF1, DF4, EF1, EM2, FM2, GF1, GM1, HF5, JF1
IV	AM5, BF1, CF2, DF2, EF5, FM5, GF4, GM3, HM1, IM4, JF4
V	AF2, BF2, DF3, EF3, EM3, FM4, GF5, GM5, HF1, HM4, JF5
VI	AF3, BM2, CF4, CM3, DM3, FM3, GF3, HF4, IF3, JF2, JM4

TABLE III

AVICAR 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
6	II, III, V	IV	I
7	I, III, IV	V	II
8	II, IV, V	I	III
9	I, III, V	II	IV
10	I, II, IV	III	V

this in mind, 160 utterances were randomly selected in each noise condition for all speaker groups (giving a total of 800 utterances per group).

B. Protocol Use

It is required to split the five groups into an adaptation set, a system tuning (i.e. evaluation) set, and a test set. In order to extend the data set (since the number of utterances in each group is limited), *k*-fold leave-one-out testing can be performed using the 5 native English groups. Averaging results over a number of folds will enable more indicative speaker-independent recognition results since individual groups may be affected by poor (or very good) performance for one or two speakers.

To facilitate adaptation, tuning and testing, 3 groups (or 60% of the data) are made available for adaptation, 1 group for tuning, with the fifth group used for testing. Ten combinations of this segregation are shown in Table III. It is intended that these experiment groupings be used in the order stated in the table. Individual studies can dictate the number of folds required.

Group VI is to be used solely as a test set since there is insufficient non-native English data to make adaptation useful, while the variation of nationalities is large enough to make system tuning problematic. The purpose of the group is to characterise the expected decrease in recognition performance when non-native English speakers use the speech recognition system under evaluation.

For all recordings, single microphone experiments are to use microphone 4 as it is centrally located in the array. Multi-microphone experiments are free to use whichever microphones are required for the particular technique.

To ensure the list of files used in each investigation is uniform, a copy of the file lists can be obtained by emailing the primary author of this paper.

IV. Recognition Experiments

In order to provide a common reference to facilitate simple comparison of results, a number of speech recognition experiments have been performed. The baseline recognition system is defined, as well as the methods for adapting clean speech models to better reflect in-car conditions. Results for all experimental folds described in Section III-B are reported in the following sections for both clean speech and adapted models.

A. Baseline Recogniser

Context-dependent 3-state triphone hidden Markov models (HMM) were trained using the 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.

In order to reflect command-and-control applications in the car environment, task grammars are chosen to be unconstrained word loops. This task grammar effectively provides the potential worst-case recognition results. For the phone number and sentences tasks, the number of words in the grammar are 11 and 773 respectively therefore constituting the small- and medium-vocabulary tasks as previously described.

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

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

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 [7].

B. Experimental Results

1) Baseline Results: Baseline results were generated using the original clean acoustic models trained as per Section IV-A. The results for the phone numbers and sentences tasks are shown in Tables IV and V respectively. Results are collated by noise condition, with the average results shown in the last column of each table being the combined accuracy over all noise conditions for a particular speaker group. The average for folds 1-5 is also shown in these tables, as are results for the non-native English speakers (i.e. Group VI). It should be noted that the results for folds 6-10 will match the results for the corresponding test group in folds 1-5. For example, fold 6 baseline results match those for fold 4 since both use speaker group I as the test set.

TABLE IV

Baseline results for the Phone Numbers task of the AVICAR database.

	Word Accuracy (%)					
Fold	IDL	35U	35D	55U	55D	Average
1	75.44	48.69	39.12	50.19	26.56	48.00
2	73.25	51.12	36.94	42.56	28.00	46.38
3	61.94	44.56	30.69	36.56	17.00	38.15
4	68.88	42.81	32.50	34.38	24.69	40.65
5	78.50	60.62	46.62	51.00	27.25	52.80
Aver. 1-5	71.60	49.56	37.18	42.94	24.70	45.20
Non-native	67.31	47.75	21.81	40.06	14.69	38.33

TABLE V

Baseline results for the Sentences task of the AVICAR database.

	Word Accuracy (%)					
Fold	IDL	35U	35D	55U	55D	Average
1	34.74	8.10	5.93	8.04	1.15	11.58
2	28.89	14.57	8.62	7.25	2.94	12.47
3	25.95	7.71	1.86	5.18	2.13	8.48
4	22.68	8.63	5.57	7.24	2.98	9.38
5	38.69	12.33	6.81	6.54	2.14	13.29
Aver. 1-5	30.23	10.25	5.75	6.85	2.26	11.04
Non-native	14.38	5.41	0.60	2.65	1.18	4.87

Analysing the results contained in these tables, a number of observations can be made. The most important of these are related to the in-car noise conditions. Comparing the results for both speeds with windows up, it can be seen that an increase in speed causes degradation in the recognition accuracy in most cases. The average decrease in performance is 7%. A similar trend is shown for the two speeds under the windows down condition.

Having the windows open appears to have more affect on the recognition accuracy than simply increasing the vehicle speed. This is demonstrated through recognition accuracies in both tasks showing better performance for the car travelling at 55mph with windows up (55U) compared to 35mph with windows down (35D). This result is in accordance with the findings of Zhang and Hansen [8] who determined that road and wind noise dominate the noise field when the windows are open. With windows open, greater decreases in accuracy occur as the speed increases (compared to windows up). This is due to increases in road and wind friction as vehicle speed increases.

The sentence task exhibits very poor performance with most conditions failing to reach an average of 10% word accuracy. The small-vocabulary task on the other hand performs much better with all noise conditions averaging above 22% accuracy. This shows considerable improvement is required to achieve both small- and medium-vocabulary speech recognition in the car environment.

A range of performance can be seen between the different speaker groups. This is particularly true when comparing folds 3 and 5 – the two folds differ by an average of approximately 15% for the phone numbers task. This observation further emphasises the need for circular experiments in order to average out the recognition accuracies to provide better indication of true speaker-independent recognition performance.

It can also be seen that the non-native speakers (group VI) perform worse than the native English speakers under all noise conditions. The performance difference is particularly noticeable when the windows are open for the phone numbers task, and under all conditions for the sentences task. This shows potential for group VI to provide the low-accuracy bound for any recognition system.

2) Adaptation Results: To test the effectiveness of the data contained in the AVICAR database for adapting clean speech models, maximum *a posteriori* adaptation (MAP) [9] was chosen. The pre-trained triphone models described in Section IV-A were assumed to give a good initial indication of the parameter distribution required by MAP adaptation. Various combinations of mean and variance adaptation were tested, as well as the amount of influence placed on the prior model (governed by the factor τ). The larger the value of τ the greater the influence placed on the model (e.g. a value of τ = 16 means the prior model has 16 times more influence than the adaptation data).

Table VI shows averaged results on the first 5 folds of the protocol for the adaptation experiments.

This table shows the advantage of performing both mean and variance adaptation, with all values of τ showing 1-2% improvement across all noise conditions. This improvement

TABLE VI

WORD ACCURACIES FOR DIFFERENT COMBINATIONS OF τ AND STATISTICS ADAPTED.

			Word Accuracy (%)					
Mean	Variance	τ	IDL	35U	35D	55U	55D	Average
Yes	No	4	81.00	75.76	67.11	72.99	54.38	70.25
Yes	Yes	4	81.34	75.54	67.08	74.13	57.60	71.14
Yes	No	8	81.14	75.81	67.58	73.60	55.05	70.64
Yes	Yes	8	81.95	76.04	68.00	74.78	58.05	71.76
Yes	No	16	81.78	75.96	67.63	73.93	55.56	70.97
Yes	Yes	16	82.44	76.79	68.85	75.76	59.06	72.58

is particularly noticeable in the 55mph with windows down condition where improvements range from 2-3%. It appears that modifying Gaussian mixture shapes through variance adaptation is important to successfully adapt clean speech models to in-car conditions.

Placing higher weighting on the prior model (i.e. larger values of τ) also makes considerable difference to the effectiveness of the adaptation. The clean speech models were trained with data from a very large number of speakers to avoid any speaker-dependency. The adaptation set has considerably less speakers (33 speakers per fold), therefore placing more emphasis on the prior model ensures the models don't become reliant on the new speakers.

Using the results from Table VI, mean and variance adaptation was performed using τ = 16 for all 10 folds of the protocol. Results for each fold can be found in Tables VII and VIII for the phone number and sentence tasks respectively. Average results have been collated for folds 1-5 and 6-10. The results shown for non-native speakers are averaged using all the adapted model sets from the first 5 folds.

The adaptation results show uniform improvements in word accuracy over the baseline results in Tables IV and V for all native English speakers. The results for the non-native English speakers are not as large since examples of such speakers are not included in the adaptation sets. For both tasks, however, there are still gains in word accuracy.

Word accuracy for the small-vocabulary (i.e. phone numbers) task shows greater improvement over the baseline results than the medium-vocabulary task. In most cases there are examples of the same speaking script in the adaptation and test sets therefore percentage model coverage for both tasks should be approximately equal. The difference is a greater number of instances of each triphone model used in the phone numbers task which can be accredited to the nature of the task – there are only 11 words, and there are 10 words per utterance. This means that models to be adapted are found more frequently than those in the sentence task. The effect is greater adaptation of the original clean speech models, resulting in better test results.

Comparing the averaged results of folds 1-5 and 6-10 for both tasks shows very minor variations in recognition accuracy. The largest variation in recognition performance between any two averages is only 0.70%. These minor variations in accuracy show reliable performance can be obtained using the leave-one-out experiment design proposed in this paper, making it suitable as a common experimental framework.

V. CONCLUSION

A continuous, speaker-independent speech recognition protocol has been proposed for the AVICAR database which enables separate investigations to make simple comparisons of in-car speech recognition results. Under this protocol, single- and multi-microphone speech enhancement techniques can be applied to the same data set. The framework also includes both small- and medium-vocabulary speech recognition.

A number of speaker groups were designed to test in-car speech recognition performance of native and non-native English speakers. To facilitate adaptation, evaluation and testing amongst these groups, a leave-one-out experiment was defined.

Baseline recognition experiments under a range of noise levels and conditions have shown general recognition rate trends which agree with previous research. In particular, opening a window in the front of the vehicle causes greater degradation in performance than simply increasing vehicle speed.

Maximum *a posteriori* adaptation using the proposed framework showed consistent accuracy improvement over baseline results for all evaluations. More importantly, using identical test data on models adapted with different data showed variations in recognition rates which were not excessive.

The observations made in this paper show that the division of speaker groups into a *k*-fold evaluation scheme provides reliable performance indicators for in-car speech recognition. The protocol is therefore suitable to be used as a common platform upon which various research efforts can be compared.

REFERENCES

[1] 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 *INTERSPEECH*, Jeju Island, Korea, 2004, pp. 2489–2492.

[2] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Automatic Speech Recognition: Challenges for the new Millenium*, Paris, France, September 2000.

[3] R. Leonard, "A database for speaker independent digit recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 1984.

[4] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognisers," *J. Acoustical Society of America*, vol. 93, pp. 510–524, January 1993.

TABLE VII

MAP adaptation results for the Phone Numbers task of the AVICAR database.

	Word Accuracy (%)					
Fold	IDL	35U	35D	55U	55D	Average
1	83.62	74.44	70.56	82.44	60.94	74.40
2	81.50	77.62	65.25	76.00	69.50	73.97
3	78.56	71.00	62.12	64.94	50.38	65.40
4	84.75	78.75	68.81	74.31	54.50	72.22
5	83.75	82.12	77.50	81.12	60.00	76.90
Aver. 1-5	82.44	76.79	68.85	75.76	59.06	72.58
6	85.06	80.75	69.69	74.94	55.56	73.20
7	81.56	76.38	65.12	77.25	68.38	73.74
8	84.06	82.88	77.94	81.81	60.44	77.43
9	77.88	72.19	63.94	65.75	49.31	65.81
10	84.19	72.69	69.25	81.06	59.62	72.69
Aver. 6-10	82.55	76.16	69.19	76.16	58.66	72.71
Non-native	72.34	66.46	50.94	73.28	42.54	61.11

TABLE VIII

MAP adaptation results for the Sentences task of the AVICAR database.

	Word Accuracy (%)					
Fold	IDL	35U	35D	55U	55D	Average
1	45.33	14.76	10.71	12.69	3.17	17.32
2	32.00	23.39	8.81	15.18	4.31	16.75
3	33.73	15.90	7.42	11.53	2.03	14.01
4	35.60	16.37	10.45	11.74	6.15	15.99
5	51.81	29.94	18.95	18.65	3.40	24.52
Aver. 1-5	39.73	20.04	11.25	13.96	3.81	17.72
6	35.70	17.46	10.64	11.35	6.44	16.25
7	31.91	24.68	9.70	15.57	4.02	17.18
8	49.76	28.18	16.58	19.82	3.88	23.62
9	36.43	17.95	9.18	8.41	2.71	14.82
10	44.27	13.14	10.13	11.14	4.13	16.55
Aver. 6-10	39.64	20.24	11.23	13.26	4.24	17.68
Non-native	17.98	8.48	1.26	2.61	0.47	6.19

[5] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, “Speechdat-car: a large speech database for automotive environments,” in *Int. Conf. on Language Resources and Evaluation*, 2000.

[6] J. Hansen, J. Plucienkowski, S. Gallant, B. Pellom, and W. Ward, “Cu-move: Robust speech processing for in-vehicle speech systems,” in *6th Int. Conf. on Spoken Language Processing*, vol. 1, Beijing, China, 2000, pp. 524–527.

[7] 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*, 3rd ed., Cambridge University Engineering Department, December 2006.

[8] X. Zhang and J. Hansen, “Csa-bf: a constrained switched adaptive beamformer for speech enhancement and recognition in real car environments,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 733–745, 2003.

[9] C. Lee and J. Gauvain, “Bayesian adaptive learning and map estimation of hmm,” in *Automatic speech and speaker recognition : Advanced topics*. Boston, Massachusetts, USA: Kluwer Academic Publishers, 1996, pp. 83–107.