LOW-COST HARDWARE SPEECH ENHANCEMENT FOR IMPROVED SPEECH RECOGNITION IN AUTOMOTIVE ENVIRONMENTS

J. Whittington, H. Ye, K. Kamalakannan, N. V. Vu,

Department of Electronic Engineering, La Trobe University, Melbourne, Australia

M. Mason, T. Kleinschmidt, S. Sridharan,

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

ABSTRACT

Voice recognition is one of the key enablers to reduce driver distraction as in-vehicle systems become more and more complex. With the integration of voice recognition in vehicles, safety and usability are improved as the driver's eyes and hands are not required to operate system controls. Whilst speaker independent voice recognition is well developed, performance in high noise environments (e.g. vehicles) is still limited. La Trobe University and Queensland University of Technology have developed a low-cost hardware-based speech enhancement system for automotive environments based on spectral subtraction and delay–sum beamforming techniques. The enhancement algorithms have been optimised using authentic Australian English collected under typical driving conditions. Performance tests conducted using speech data collected under variety of vehicle noise conditions demonstrate a word recognition rate improvement in the order of 10% or more under the noisiest conditions. Currently developed to a proof of concept stage there is potential for even greater performance improvement.

INTRODUCTION

As in-vehicle systems such as navigation, non critical system control (e.g. HVAC, cruise control) and entertainment devices become more and more complex, automatic speech recognition (ASR) is potentially a key technology for reducing driver distraction. Speech-based systems offer a potential increase in the safety and usability of in-car devices as the driver's eyes and hands are not required to operate system controls. Despite speech recognition being well developed, performance in high noise environments – such as those encountered in vehicles – is still well below consumer expectation. Which even for the operation of non-critical systems needs to be accurate more than 95% of the time. Speech enhancement techniques are a way of improving the performance of these systems in the full range of noise conditions.

Currently, research into in-car speech recognition systems has been confined to the United States of America, Europe and Asia. No significant and publicly available speech data exists for Australian-accented speakers under Australian driving conditions. Thus current voice recognition systems are less suited to the Australian accent, resulting in further reductions in recognition performance. This lack of appropriate data restricts research into developing and analysing systems to suit the Australian market.

The lack of Australian in-car speech data and the general poor performance of in-car speech recognition systems has motivated research at the Queensland University of Technology and La Trobe University, suported by the AutoCRC, into speech enhancement techniques for noise reduction as well as the collection of data suitable for analysing performance under Australian conditions. This research developed a number of techniques suitable for noise reduction and incorporation with in-car speech recognition systems. These techniques include single-microphone techniques currently most suited to vehicle applications through reduced production costs, as well as two-channel techniques which have the potential to provide superior noise reduction performance.

Developing theoretical models for enhancement techniques is only half the challenge to widespread deployment in the automotive industry. Therefore, the proposed enhancement techniques were developed into low-cost hardware implementations for real-time operation. These implementations were made sufficiently cost-effective that further development and integration with other in-car systems is possible in order to keep production costs to a minimum.

SPEECH ENHANCEMENT

Speech enhancement algorithms are a family of techniques designed to make speech recognisers more robust by reducing the levels of noise present in speech signals, thereby enabling the use of clean speech models for which data is readily available to train effective models. This is a excellent approach for the provision of robust speech recognition as little-to-no prior knowledge of the operating environment is required for improvements in speech recognition accuracy.

Speech enhancement algorithms typically perform their primary processing in the frequency domain. This is true of both enhancement techniques discussed in this paper (Spectral subtraction and delay-sum beamforming). The general approach to processing a speech signal in the frequency domain is presented in Figure 1. While it is possible to avoid the overlap-add reconstruction by incorporating the enhancement into the speech recogniser front-end, this work uses the reconstruction method in order to interface with existing speech recognition engines.

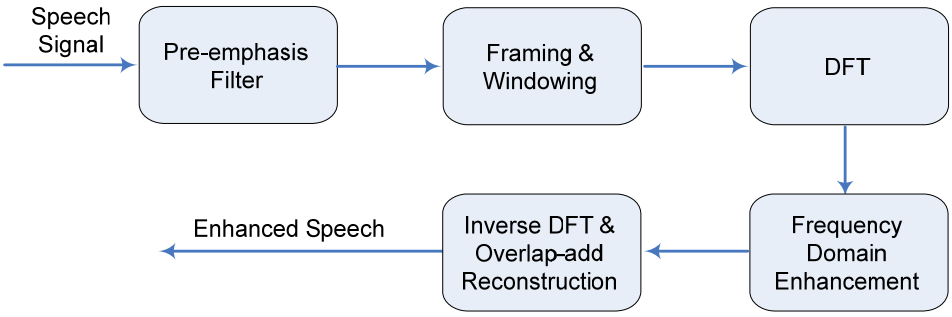


Figure 1: Diagram of basic speech processing and enhancement in the frequency domain.

After a speech signal is acquired via a microphone, it is passed through a pre-emphasis filter which ensures a flatter signal spectrum by boosting the amplitude of the higher frequency components relative to the lower frequencies. The signal is then decomposed into a series of frames of a set length (typically 32ms – at 16 kHz this represents a frame length of 512 samples) and a Hamming window is applied to each frame in order to attenuate discontinuities at the frame edges. The frames are created using a sliding window with frame advances typically being 50% the length of the frame.

A Discrete Fourier Transform (DFT) then transforms the time-domain acoustic waveform to a discrete frequency representation. The enhancement process operates on the frequency domain representation, altering each frame's spectrum in an effort to improve the signal. Following frequency domain enhancement, each frame is transformed back to the time domain using an inverse DFT and adjacent frames are resynthesised using an overlap-add technique to produce an enhanced time-domain signal. The enhanced signal can then be used for as an input for automatic speech recognition.

Spectral Subtraction

In a noisy environment, clear speech *s(n)* is impacted by predominantly uncorrelated, additive background noise *d(n)* to produce corrupted speech *y(n)* as follows:

$$y(n) = s(n) + d(n) \qquad \text{Eq. 1}$$

Obviously, if we are able to produce a good estimate of the background noise *d(n)* and subtract this from the corrupted speech signal *y(n)* we should be able to generate a good approximation of the original speech signal *s(n)*. To perform this effectively in the time domain a continuous noise estimate, unaffected by the speech, is needed. While this is possible for some related applications, such as, noise cancelling headphones, it is not viable in real time when the speech and noise are being picked up by the same microphone. In this case any significant level of speech signal appearing in the noise estimate would result in a corresponding reduction in the desired speech signal, at best, negating the speech enhancement benefits and at worst, further corrupting the wanted speech signal.

An alternative is to operate the process in the frequency domain, where equation 1 becomes:

$$Y(i,\omega) = S(i,\omega) + D(i,\omega) \qquad \text{Eq. 2}$$

Here, $S(i,\omega)$ is the discrete frequency representation of the original (wanted) speech signal, with $D(i,\omega)$ the discrete frequency representation of the noise, and $Y(i,\omega)$ the discrete frequency representation of the corrupted speech. If we can make the following assumptions: (a) that the utterances for speech recognition are relatively short, that is, typically a word or few words; (b) that there is sufficient non speech time to generate an accurate noise estimate; and (c) that the noise estimate is sufficiently accurate to remain valid during the utterance and does not require updating while speech is taking place; then real time speech enhancement is possible. As these assumptions are generally valid for voice control applications in an automotive environment, the spectral subtraction technique can be used to improve in-car speech recognition. This description outlines the essence of the spectral subtraction technique where spectra of the noise estimate is subtracted from the spectra of the corrupted speech to give the approximate spectra of the original speech. A more in-depth analysis can be found in the following references (Berouti et al. 1979)(Boll 1979)(Martin 1994)(Kleinschmidt 2010).

Delay-Sum Beamforming

Beamforming is a method of spatial filtering that differentiates the desired signals from noise and interference according to their locations. The direction where the microphone array is steered is called the look direction. One beamforming technique is the delay-and-sum beamformer which works by compensating signal delay and attenuation to each microphone appropriately before they are combined using an additive operation. The outcome of this delayed signal summation is a reinforced version of the desired signal and reduced noise due to destructive interference among noises from different channels.

Beamforming speech enhancement techniques can be implemented in one of two domains – near field or far field. In the near field, the sound source (i.e. the speaker) is assumed to be in close proximity to the sensor (i.e. the microphones). This is an appropriate assumption for in-car environments as the speaker is typically less than 1m from the microphones. In the near field, the wave front is assumed to be spherical since the sound source and sensor are in close proximity. This supposition allows calculation of accurate time delay estimation and proper attenuation as a function of speaker distance from each microphone in the array. The time delay to any microphone is calculated with respect to a reference microphone:

$$\tau_n = \frac{d_n - d_{ref}}{c} \qquad \text{Eq. 3}$$

where *c* is the speed of sound (~340m/sec), d_n and d_{ref} are the distances from the source to the n^{th} and reference microphones respectively. Given the time delay in Equation 3, the resulting output signal (in the time-domain) can be formulated as:

$$y(t) = \frac{1}{2}\sum_{n=1}^{2} \frac{d_{ref}}{d_n} x_n(t - \tau_n) \qquad \text{Eq. 4}$$

where x_n is the signal incident on the n^{th} microphone. This equation can be represented by the delay and gain blocks shown in Figure 2. While in general the greater the number of elements

the more accurate the beamformer, we have implemented a dual microphone system as this is most appropriate in terms of minimising production costs in the automotive industry. The above provides a basic overview, greater detail of delay-sum beamforming and our implementation can be found in the following references (Ye et al. 2009)(Johnson & Dudgeon 1992)

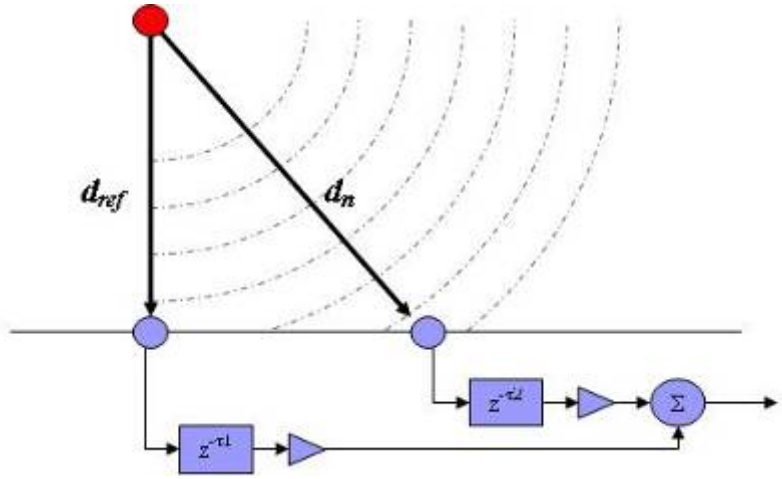


Figure 2: **Near field block formulation of delay-sum beamforming.**

DEVELOPMENT OF REAL-TIME HARDWARE

Initially the two speech enhancement techniques were developed as MATLAB scripts using high precision, complex floating-point arithmetic. MATLAB is an excellent platform for research, algorithmic exploration, and determination of appropriate parameters for effective in-car speech enhancement performance to improve speech recognition rates. However, it does not run in real-time and requires a processing platform of performance and cost that is unsuited to wide spread adoption in the automotive environment.

To achieve this goal the speech enhancement algorithms require adaption for operation on a realisable hardware platform suitable for use in an automotive environment. The hardware must be cost-effective to meet the price sensitivities of automotive manufactures, while providing real-time performance and considerable Digital Signal Processing (DSP) capacity. Low-cost field programmable gate arrays (FPGAs) especially manufactured for automotive applications are an excellent candidate for this application.

Justification for FPGA-Based Implementation

The majority of current automotive electronics are powered by low-cost embedded processors that perform multiple tasks including CAN network communications and HMI. Currently only a modest amount of automotive electronics are based on FPGAs, primarily due to their higher single-unit cost compared to an embedded processor. The market is changing since the cost deferential is insignificant considering the much higher performance of an FPGA. This performance is coupled with the fact that even modest-sized FPGAs may contain multiple instantiations of embedded processors as well as other specialised hardware elements, such as, a speech processing and enhancement system. This eliminates the need for multiple devices, simplifying overall design and cost. Recognising this market opportunity, Xilinx, a leading FPGA vendor, has developed the Xilinx Automotive (XA) product family specifically for automotive applications (Kitagawa, K 2007).

With this in mind, Xilinx devices and development tools were chosen for this work as a clear pathway to a commercialisable platform is available. Speech enhancement algorithms rely on considerable DSP power. Since cost is a key factor to eventual widespread adoption in the automotive field, target devices must be cost effective while still providing relatively high performance DSP. With well over one million system gates, plus memory and XtremeDSP^{TM} slices, Xilinx XA Spartan-3A DSP FPGAs fit this requirement well. As a final target for this work the Spartan-3A DSP 1800A FPGA was chosen. Being a general production equivalent to its

Xilinx Automotive cousin, successful implementation in one device demonstrates capability for implementation on the other.

The Spartan-3A DSP devices are lower cost member of the Xilinx XtremeDSP™ portfolio, which also contains the larger and higher performance Virtex-4 SX and Virtex-5 SXT. Due to similarities in their architecture - particularly the XtremeDSP™ slices - designs can be transported between XtremeDSP™ devices in a reasonably straight-forward manner. This feature enables designs to be initially developed in a high-end device and gradually reworked towards a lower-end device solution. (Kitagawa, K 2007)(Bagni & Zoratti 2007).

FPGA Design Process Overview

Moving from an algorithmic description to a quality, cost effective FPGA solution is anything but trivial. The spectral subtraction and delay-sum algorithms were originally developed as MATLAB scripts using high precision, complex floating-point arithmetic. A one-to-one conversion to an equivalent FPGA implementation cannot be achieved as many of the complex operations cannot be directly or easily implemented in an FPGA with reasonable (ideally minimal) resource utilisation. Options for the implementation of such operations involve use of approximations, either formulae-based or through the use of look up tables, both of which can introduce error to the system. Also, the precision of data in the FPGA implementation is limited to a fixed number of bits (fixed-point representation) which results in the addition of quantisation noise to the system. This necessitates a multi-step process with considerable testing at each stage, summarised broadly as follows.

1. Conversion of floating-point model to a fixed-point (data and operations) implementation in MATLAB, mirroring the major blocks expected in the FPGA implementation.
2. Comprehensive testing of the fixed-point MATLAB design against the floating-point version, block-by-block and at complete system level. Adjustment of fixed-point model as required.
3. Implementation of the fixed-point design as Xilinx System Generator™ (XSG) models. XSG is an FPGA hardware DSP development environment that sits above MATLAB and Simulink software packages.
4. Comprehensive testing of each major block of the XSG design against its fixed-point MATLAB equivalent, and testing of the complete XSG model against both the fixed-point and floating-point MATLAB versions. Adjustment of design as required.
5. Generation of hardware description language (HDL) design from the completed XSG model, followed by synthesis using Xilinx tools, and implemention on a high-end FPGA.
6. Check of FPGA resource usage of the design, with analysis, block-by-block, to identify resource inefficiencies. Optimise design to use more appropriate resources.
7. Undertake speech recognition tests comparing FPGA performance against floating-point algorithm. Adjust/optimise design as appropriate.
8. Implement design on the target low-end Spartan-3A DSP FPGA. Test output sample-by-sample against high-end version, using a variety of dedicated test input waveforms and various speech samples. Repeat steps 6 & 7.

Further details of the FPGA design process can be found in the following references (Whittington et al. 2008)(Whittington et al. 2009)(Ye et al. 2009).

Resource Usage

Each FPGA device contains a fixed amount of various resources that can be used to implement designs. Resources not used by one particular design are potentially free for use by other (one or more) designs which can operate independently on the same device. This is somewhat analogous to two or more dwellings built on a single block of land. Table 1 shows key FPGA resource usage for the final low-cost hardware implementations of the Spectral Subtraction and Delay-Sum Beamforming designs.

Table 1: Summary of resource utilisation for Spectral Subtraction and Delay-Sum FPGA designs in Spartan-3A DSP 1800A.

	Spectral Sub			Delay-Sum		
Resource Type	**Available**	**Used**	**Usage (%)**	**Available**	**Used**	**Usage (%)**
Slices	16640	2196	13.2	16640	1621	9.5
BRAM	84	10	11.9	84	16	19
DCM	8	1	12.5	8	1	12.5
DSP48	84	25	29.8	84	22	26

The Spectral Subtraction design uses just over 13% of the total (general FPGA logic fabric) slices available, and nearly 30% of the DSP48 XtremeDSP^{TM} blocks. The larger percentage use of the DSP48 blocks is expected due to the intensive DSP requirements of the algorithm. The percentage use of other key resources, Block-RAM (BRAM) and digital clock manager (DCM) blocks are of a similar level to the slice usage. By comparison the Delay-Sum design at 9.5% uses less slices, but more BRAM at 19%, this is due to significant buffering of the dual channel input to enable valuable processing resources to be shared between the channels. Finally, DCM usage is the same, while the use of DSP blocks is slightly less.

Overall, the each design uses around 1/7th of the FPGA resources available in the Spartan-3A DSP 1800 device apart from the specialised DSP48 blocks of which more than 70% remain free for other uses. This low resource utilisation enables other processes (such as CAN communications or HMI providing infotainment, driver information and driver assistance) to be incorporated into a single FPGA. By amortising implementation costs over a number of applications, overall manufacturing component costs would be kept to a minimum.

VALIDATION OF PERFORMANCE

Having constructed the two speech enhancement systems it is necessary to validate their performance. For this work there are two aspects of interest: (i) to what extent do the two speech enhancement techniques improve speech recognition performance; and (ii) how well does the fixed-point FPGA hardware implementation match the floating-point software implementation.

Validation of speech enhancement performance in this context can only be measured through statistical analysis of speech recognition rates for various enhancement scenarios, including the no enhancement case, using large data sets containing a variety of speakers. Experiments for this work were conducted using the the Australian In-Car Speech and the AVICAR databases.

Speech recognition experiments involved passing large sets of speech waveforms from these databases through each of the various enhancement implementations in turn, floating point and FPGA for both spectral subtraction and delay and sum beamforming. Each set of enhanced files were subsequently passed through a speech recognition engine and word accuracy statistics collected. To provide a baseline measure the data sets were also passed directly through the speech recognition engine (no enhancement case).

The Australian-English in-car Speech Database

The Australian-English In-Car Speech (AEICS) Database is a multi-channel recording of a series of prompts from an in-car navigation task collected over a range of Australian speakers in a variety of live driving scenarios common to Australian driving conditions. This task reflects the primary application of the user demographic most likely to benefit from in-car speech recognition interfaces – professional drivers.

The purpose of the database is to provide a rich resource of speech data appropriate for investigating speech processing needs in the adverse environment of the car cockpit. The multi-channel recording process provides the capability for evaluating the performance of speech enhancement techniques proposed as part of this project for improving in-car speech recognition accuracies. The task oriented grammar of the database also provides the potential to investigate language processing techniques which may aid in medium vocabulary command-and-control applications.

Recording Environment

The database was collected in a 2008 VE Commodore specifically outfitted with eight high-quality Sennheiser microphones, an in-car PC and LabView software used to record the data. The microphones were fitted to the central roof console pointing downwards as shown in Figure 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. The microphones labelled M0 (closest to driver) to M7 (closest to passenger) were spaced symmetrically around the midline of the vehicle with 2cm spacing between each adjacent microphone. The average location of the driver's mouth was estimated (with reference to microphone M0) to be 35cm to the right, 25cm below, and 17.5cm behind this reference microphone (i.e. ~46.4cm in a direct line).

Figure 1: **Position of microphone array inside car cabin.**

A total of 50 native English speakers were recorded for the database with 20 female speakers and 30 male speakers represented. The recordings were focused around a mock navigation task with typical comands used to control a in-car navigations system, including, various Australian suburbs, street names, prefixes and types along with numbers, and various directives (e.g. enter, start stop etc.).

Utterance recording was conducted under seven different driving conditions. These conditions were chosen to capture a variety of general noise types and levels present in the cabin of a vehicle whilst also representing likely driving scenarios in Australia with the expectation being that these conditions will provide enough information to predict performance variation with changes in background noise conditions. Table 2 shows a full list of the recording conditions. The acronym HVAC stands for Heating, Ventilating, and Air Conditioning system. Full details of The Australian In-Car Speech Database can be found in the research paper "The Australian English Speech Corpus for In-Car Speech Recognition" (Kleinschmidt, et al. 2009)

Table 2: In-car noise conditions present in the Australian In-Car Speech Database

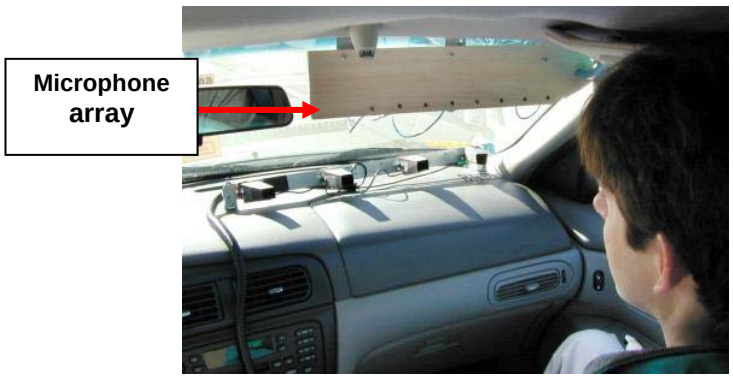
Condition	Description
C0	Car idle, sealed cabin, no HVAC
C1	Medium speed (50-60 km/h), sealed cabin, no HVAC
C2	Medium speed (50-60 km/h), sealed cabin, HVAC on high fan

C3	Medium speed (50-60 km/h), driver's 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 high fan
C6	Car idle, sealed cabin, HVAC on high fan

AVICAR Database

AVICAR is a multi-channel Audio-Visual In-CAR speech database collected by the University of Illinois, USA. It is a large, publicly available speech corpus designed to enable low-SNR speech recognition through combining multi-channel audio and visual speech recognition. For this collection, an array of eight microphones was mounted on the sun visor with four video cameras mounted on the dashboard in front of the speaker who was positioned on the passenger's side of the car, as shown in figure 2. The location of the speaker's mouth was estimated to be 50 cm behind, 30 cm below and horizontally aligned with the fourth microphone of the array (i.e. ~58.3cm in a direct line). The microphones in the array are spaced 2.5 cm apart.

Figure 2: **Position of microphone array inside car cabin AVICAR (UIUC 2006).**



Utterances for each speaker were recorded under five different noise conditions which are outlined in Table 3. Four different speaking tasks were targeted during this collection – isolated digits, isolated letters, phone numbers (i.e. digit sequences) and sentences (from the TIMIT standard) (Lee, et al. 2004)(UIUC 2006). For the evaluations in this work, only the continuous speech phone numbers task is utilised.

Table 3: AVICAR database in-car noise conditions

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

For this work, a continuous speech recognition evaluation protocol similar to that previously outlined for the Australian In-Car Speech Database was formulated. This protocol consists of 55

speakers separated into 5 groups enabling adaptation, tuning and testing. Details of this protocol can be found in the published research paper “A Continuous Speech Recognition Evaluation Protocol for the AVICAR Database” (Kleinschmidt, et al. 2007)

Speech Recognition System

A speaker-independent speech recognition engine, operating in a similar manner to commercially available products, developed to provide a common reference for the evaluation speech enhancement performance. Key details are as follows. 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) (Young et al. 2006).

All speech recognition results quoted below are word accuracies (in %), calculated as:

$$WordAccuracy = \frac{N - D - S - I}{N} \times 100\% \qquad \text{Eq. 5}$$

Where:

- N represents the total number of words in the experiment;
- D the number of correct words omitted in the recogniser output;
- S the number of incorrect words subtituted for a correct word in the output;
- I the number of extra words added to the recogniser output.

Baseline Recognition Performance

Before measurements of speech recognition improvement through the application of enhancement techniques can be accomplished a baseline of the performance of the recognition engine for each condition in the two databases must be established. Table 4 shows the baseline speech recognition results for the AEICS database for microphone 0 (i.e. the microphone closest to the driver). Baseline test conducted similarly for the other microphones indicated a trend of decreasing word recognition accuracy as the microphone was placed further from the driver (not unexpectedly).

Table 4: Baseline speech recognition results for the AEICS database.

Noise Condition							
C0	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**Average**
84.89	69.69	34.06	53.01	53.88	30.54	41.16	52.01

Condition C0 (idle, no HVAC) can serve as the baseline result for comparison with other conditions since it exhibits the least background noise. For conditions C1 and C4 (travelling at 60 km/h and 100 km/h respectively with windows closed and HVAC off) a progressive decrease in accuracy for both baseline recognisers is observed. The main contribution to noise in these conditions is the tyre and road noise transmitted into the vehicle cabin. As expected, noise generated from higher car speeds contributes more to the degradation of recognition accuracy. The same trend can be observed when comparing C6 to C2 and C5 (i.e. same speeds but with HVAC on), however the fan noise contributes significantly to the overall background noise.

When HVAC is turned on (C6, C2 and C5), recognition accuracies drop significantly compared to the corresponding conditions without HVAC. For example, in the idle + HVAC case (C6), is

less than half the idle only case. Even the medium speed with driver's window down case (C3) performs better.

Overall, the fan noise generated from the HVAC fan is the most significant source of background noise in terms of recognition accuracy degradation. The environmental noise introduced from outside the car when the car windows are opened also affects the speech recognition performance but not as severely. While increasing speed (noticeable through tyre and road noise) also contributes to accuracy drop, it has the least effect of the different types of background noise (when windows are closed).

Table 5 shows baseline speech recognition results for the AVICAR phone numbers task. For these experiments microphone 4 was used as it is central to (and hence closest to) the speaker.

Table 5: Baseline speech recognition results for the AVICAR phone numbers task.

Noise Condition					
IDL	**35U**	**35D**	**55U**	**55D**	**Average**
71.52	49.56	37.18	42.77	24.61	45.13

Similarly to the C0 condition in the AEICS database, the IDL (Idle) Condition can be used as a baseline result for comparison with other conditions as it possesses the least background noise. While it is not possible to make direct comparisions between databases, as the recording conditions are not identical, similar trends can be observed. Again for a sealed cabin increased speed (IDL-35U-55U) results in greater road noise and a corresponding drop in accuracy. Also, when the windows are down increased noise results in a lower accuracy compared to a sealed cabin the same speed (33D & 35U, 55D & 55U). As the use of HVAC was not included in the AVICAR database no comparison can be made in this regard.

Spectral Subtraction Validation

Speech recognition experiments were performed as indicated previously on both the floating point and FPGA implementations of the spectral subtraction algorithm and compared with the baseline measurements. Table 6 shows recognition accuracies for the AEICS database, while Table 7 displays results for the AVICAR phone numbers task.

Table 6: Speech recognition results (word accuracies in %) with floating-point and FPGA implementations of spectral subtraction applied to the AEICS database.

	Noise Condition							
	C0	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**Average**
Baseline	84.89	69.68	34.06	53.01	53.88	30.54	41.16	52.01
SpecSub (Floating-Point)	86.88	76.17	48.31	60.57	61.51	45.24	52.76	61.27
SpecSub (FPGA)	86.88	76.21	48.39	59.75	61.63	45.35	52.88	61.21

Table 7: Speech recognition results (word accuracies in %) for floating-point and FPGA implementations of spectral subtraction applied to the AVICAR phone numbers task.

	Noise Condition					
	IDL	**35U**	**35D**	**55U**	**55D**	**Average**
Baseline	71.52	49.56	37.18	42.77	24.61	45.13
SpecSub (Floating-Point)	74.81	54.74	40.85	50.70	30.71	50.36
SpecSub (FPGA)	74.66	54.76	40.86	50.55	30.67	50.30

Speech enhancement performance

Both sets of results clearly show the effectiveness of spectral subtraction as an enhancement technique for in-car speech recognition. In all noise conditions across the two data sets there is an improvement in recognition accuracy. This is also true of the idle cases where the background noise levels are already very low – the minimum improvement in these conditions was almost 2%. Spectral subtraction is particularly effective in the scenarios where the HVAC system is turned on (i.e. C6, C2 and C4 in the AEICS), with improvements in the idle, 50-60 km/h and 90-100 km/h conditions being 11.6%, 14.25% and 14.7% respectively.

FPGA implementation

Analysing the results it can be seen that the fixed-point FPGA implementation of spectral subtraction performs well across the range of in-car noise scenarios. There is obvious speech recognition improvement over the baseline results in all cases and the performance matches that of the floating-point implementation to within +/- 0.15%, with the exception of the 50-60 km/h with window down condition (C3) in the AEICS database, where the difference is -0.82%. This clearly demonstrates that the current low-end FPGA design can provide effective speech enhancement in a low cost and compact form suitable for use in automotive environments.

While this is a good result so far, further investigation of the performance of the fixed-point design under the C3 condition may lead to additional improvement. For example, a subsequent spectrum analysis of many sample files from the C3 noise condition indicated very high amplitude values in the low-frequency range (compared to higher frequencies). This is most likely the result of microphone vibration due to wind (from the open window). Which in turn could lead to a loss of precision in the fix-point implementation, as internal scaling attempts to make room for very high amplitude values. An interesting artefact is that for the AEICS, apart from the C3 condition, the FPGA implementation matches or exceeds the performance of the floating-point algorithm. Should access to a wider range of speech databases with similar noise conditions be available, and the statistical difference be maintained, then further research could lead to improvements in the spectral subtraction design.

Delay-Sum Beamforming Validation

The delay-sum beamformer implementations were tested on both the AEICS and the AVICAR databases in a similar manner to the spectral subtraction implementations. One important difference is that two microphone inputs are required in this case. Although, as the delay-sum beamformer provides a single channel output the remainer of the process remains the same. For the AEICS database, microphones 0 and 3 were used, whilst the AVICAR database utilised microphones 2 and 6. It should be noted here that the AVICAR database provides an indication of the performance of the delay-sum beamformer using a symmetric array (i.e. both microphones at an equal distant from the speaker), whilst the AEICS database analyses the performance using asymmetric arrays. Table 8 shows recognition accuracies for the AEICS database, while Table 9 displays results for the AVICAR phone numbers task.

Table 8: Speech recognition results (word accuracies in %) with floating-point and FPGA implementations of delay-sum beamforming applied to the AEICS database.

	Noise Condition							
	C0	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**Average**
Baseline	84.89	69.68	34.06	53.01	53.88	30.54	41.16	52.01
Delay-Sum (Floating-Point)	84.73	69.08	40.26	42.75	52.70	36.57	47.08	52.80
Delay-Sum (FPGA)	84.65	69.25	40.63	43.01	53.31	36.97	47.85	53.16

Table 9: Speech recognition results (word accuracies in %) for floating-point and FPGA implementations of delay-sum beamforming applied to the AVICAR phone numbers task.

	Noise Condition					
	IDL	**35U**	**35D**	**55U**	**55D**	**Average**
Baseline	71.52	49.56	37.18	42.77	24.61	45.13
Delay-Sum (Floating-Point)	80.39	64.07	53.01	56.96	37.11	58.31
Delay-Sum (FPGA)	80.03	64.06	53.08	56.69	36.71	58.11

Speech enhancement performance

For the AEICS performance improvement can only be seen in conditions with the HVAC turned on (C2, C5 and C6). The remaining noise conditions all experience decreases in performance. While marginal in the C0, C1 and C4 conditions, a significant decrease is seen for condition C3, which is 50-60 km/h with driver's window down. Due to the microphones being placed to the left of the driver (and also the driver's window) the beam is pointed towards the driver and beyond the driver, the open window. It appears that beam emphasises the noise coming from outside as much (and probably more) than the target speech. A possible solution to this is to direct the beam more towards the driver's shoulder, and therefore steer the beam away from the window.

The results for the AVICAR database show impressive and global improvements in speech recognition accuracy. These range from 8.51% (IDL) to 15.9% (35D) and in all cases provide a significantly greater improvement than spectral subtraction. An advantage of the symmetric array used here could be that even if the formed beam is not quite central to the speaker's mouth elements of the beam will fall inside the vehicle on material less likely to reflect or transmit sound, such as, speaker's clothing, seat upholstery, interior lining etc. In contrast the asymmetric array used by AEICS database places elements of it's beam on the drivers's window, which even when shut it is a good material for reflecting sound and possibly even transmitting sound through vibration.

These results indicate an advantage in using a symmetric array for delay-sum beamforming in vehicle environments. The AVICAR database shows clear improvements in a range of noise conditions, whilst the AEICS database only improves in the scenarios when the HVAC system is turned on (incidentally, the majority of the fan noise comes from directly beneath the array and so the beam steering filters out these components). This leads to the conclusion that microphone arrays are best placed directly in front of the driver for in-car speech recognition, even though this means a move away from an industry-favoured microphone position.

FPGA implementation

The recognition rates for the speech data processed using the FPGA hardware implementation are comparable to those generated using the floating-point model. In fact overall for the AEICS database the FPGA implementation provides marginally better performance. On average, the FPGA implementation provides a +0.36% improvement for the conditions encountered in this database. While for the AVICAR database the floating-point algorithm typically performs slightly better than the FPGA implementation. The maximum difference observed in this database is -0.40%.

Analysing these results, it is expected that the quantisation errors in the fixed-point (FPGA) representation of the delay filters will result in a less defined location of the target speaker. While we would normally expect this to result in a reduction in performance, for the AEICS database this may place the beam slightly away from significant noise sources. Resulting in the improvement we see in this case. While another factor could be that the fixed-point beam position may actually result in a better placement when the position of the driver is different from the “average” location assumed for these experiments On the other hand, for the AVICAR database the floating-point beam is most likely well positioned, and so the small positioning error introduced by fixed-point implementation produces an overall slight degradation in performance. In general, as the FPGA hardware implementation of delay-sum beamforming provides very similar improvements in speech recognition accuracy to that of the floating point implementation, we can conclude that a suitable hardware system has been successfully implemented.

DISCUSSION

The optimised hardware algorithms for both single-channel spectral subtraction and dual-channel delay-sum beamforming have been shown to improve speech recognition performance under a range of automotive noise conditions. The spectral subtraction component is particularly adept at mitigating noise from the HVAC system, whilst the delay-sum beamformer can pinpoint the driver’s speech over noise sources in the vehicle. This could even be applied to favour the driver’s speech over other occupants of the vehicle. While the capability of each technique has been demonstrated, there is potential for further research towards even greater enhancement capability. Including, combining the two techniques into the one speech enhancement system.

Each design fits easily into a Spartan-3A DSP 1800 FPGA with around 85% of the general FPGA resources and more than 70% of the specialised DSP48 blocks remaining free for use by other applications. This FPGA is a general production equivalent of a low-cost Xilinx automotive device, thus the designs can be readily ported into automotive rated hardware. Furthermore, as these speech enhancement techniques only occupy a modest portion of the FPGA, other systems could share the one hardware platform, reducing overall components and costs.

As part of the validation process the unique Australian-English In-Car Speech Database has been created under Australian driving conditions with Australian accented speakers. Not only has this speech corpus proved useful for optimising and testing the enhancement algorithms for use in Australian vehicles, some of the parameters used in this data collection have lead to interesting findings. The “industry preferred” microphone placement used (on the central roof console pointing downwards) proved very adept a picking up noise from an active HVAC system and when on high, this noise was the dominant factor impacting on recognition performance. When delay-sum beamforming was used, the central microphone location required the beam to be directed in part towards a window, impacting performance, particularly when open. This contrasted markedly to the very good speech enhancement performance when microphones placed to the front of the speaker are used. The conclusion drawn is that microphone placement is a key factor in vehicular speech enhancement and recognition performance.

While it has been clearly demonstrated that cost effective speech enhancement hardware can be produced for automotive applications, speaker independent recognition rates under adverse conditions are still well below that required for general consumer acceptance, even for the control of non-critical functions, such as, the radio or seat adjustment. Thus, considerable work is still required to improve speech enhancement technology and the speaker independent performance of speech recognition systems (which is beyond the scope of this work). Though it may eventually come to pass that speech recognition is used for the control of more critical

automotive functions this would require a vast improvement in performance along with the integration of back-up fail-safe mechanisms.

Beyond the automotive environment, this speech enhancement system has potential to improve speech recognition performance in any noisy environment. Other applications which could utilise this technology include: voice control in manufacturing environments for part selection, voice operated consumer information kiosks, such as a public transport information portal located at a bus stop, control of home appliances and robots by voice, speech control of wheel chairs as an aid for the handicapped.

CONCLUSION

Noise reduction techniques have been studied and optimised, using authentic Australian English collected under typical driving conditions, for incorporation with in-car speech recognition systems. These have been shown to provide significant improvement, in the order of 10% or more under the noisiest conditions, over use of recognition engine alone. Performance evaluations have provided an insight into the behaviour of these techniques in particular with differing microphone configurations and noise conditions. There remains considerable scope for further research on in-car speech enhancement hardware, including the impact on microphone deployment on noise pickup and speech enhancement capability.

Overall, this research demonstrates that cost-effective real-time hardware implementations of speech enhancement techniques can significantly improve the modest performance of speech recognition engines in high noise environments, such as, vehicles. While considerable work is still required to increase recognition rates to an acceptable level, these improvements point the way to the realisation of in-car voice control for non-critical applications as standard in vehicles of the future. This application not only promises greater user convenience, but a significant safety benefit through reduced driver distraction.

ACKNOWLEDGMENTS

This work was supported in part by the Australian Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

REFERENCES

Bagni, D and Zoratti, P 2007, “Block matching for automotive applications on Spartan-3A DSP devices,” *XCell Journal*, no. 63, pp. 16–19, 2007.

Berouti, M, Schwartz, R, and Makhoul, J, 1979, “Enhancement of speech corrupted by acoustic noise,” *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1979, Washington D.C. USA*, 1979 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings, pp 208–211.

Boll, SF 1979, “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Trans. Acoust. Speech and Signal Processing*, vol ASSP-27, pp. 113-120, April 1979.

Johnson, DH and Dudgeon, DE, 1992, *Array Signal Processing: Concepts and Techniques*. Simon & Schuster, 1992.

Kitagawa, K 2007, “At the heart of consumer and automotive innovation,” *XCell Journal*, no. 63, pp. 12–13, 2007.

Kleinschmidt, T, Dean, D, Sridharan, S, and Mason, M 2007, “A continuous speech recognition protocol for the AVICAR database”, *1st International Conference on Signal Processing and Communication Systems, Gold Coast, QLD, Australia, 2007,* Proceedings of the 1st International Conference on Signal Processing and Communication Systems, pp. 339–344.

Kleinschmidt, T, Mason, M, Wong, E, and Sridharan, S 2009, “The Australian English Speech Corpus for In-Car Speech Processing,” *IEEE International Conference on Acoustics, Speech,*

and Signal Processing, 19-24 April, 2009 Taipei, Taiwan, 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings, pp 4177-4180.

Kleinschmidt, T 2010, "Robust Speech Recognition using Speech Enhancement", *PhD dissertation*, School of Engineering Systems, Queensland University of Technology, March 2010

Lee, B, Hasegawa-Johnson, M, Goudeseune, C, Kamdar, S, Borys, S, Liu, M, and Huang, T, 2004, "AVICAR: Audio-visual speech corpus in a car environment," *ICSLP 8th International Conference on Spoken Language Processing* Jeju Island*,* Korea *October 4-8*, *2004,* INTERSPEECH 2004, pp. 2489–2492.

Martin, R 1994, "Spectral subtraction based on minimum statistics," *VII European Signal Processing Conference, September 13-16, 1994, Edinburgh, Scotland*, *U.K.*, Proceedings of EUSIPCO 1994, VII European Signal Processing Conference, pp. 1182–1185.

University of Illinois at Urbana Champaign (UIUC) 2006, *AVICAR Project: Audio-Visual Speech Recognition at UIUC*, viewed 26 April 2010, www.ifp.uiuc.edu/speech/AVICAR/

Whittington, J, Deo, K, Kleinschmidt, T 2008, and Mason, M., "FPGA Implementation of Spectral Subtraction for In-Car Speech Enhancement and Recognition," *2nd International Conference on Signal Processing and Communication Systems, 15-17 December, 2008, Gold Coast, QLD, Australia,* Proceedings of the 2nd International Conference on Signal Processing and Communication Systems, pp 1-8, ISBN 978-1-4244-4242-3.

Whittington, J, Deo, K, Kleinschmidt, T, and Mason, M 2009, "FPGA Implementation of Spectral Subtraction for Automotive Speech Recognition," *IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems, 30 March-2 April, 2009, Nashville, TN, USA,* 2009 IEEE Symposium on Computational Intelligence in Vehicles and Vehicular Systems (CIVVS 2009) Proceedings, pp. 72-79, ISBN 978-1-4244-2770-3.

Ye, H, Whittington, J, Himawan, I, Kleinschmidt, T, and Mason, M 2009, "FPGA Implementation of Dual-Microphone Delay-and-Sum Beamforming for In-Car Speech Enhancement and Recognition," *AutoCRC Conference 2009, 5 March, 2009, Melbourne, VIC, Australia,* AutoCRC Conference 2009 Smarter, Safer, Cleaner – Proceedings, pp 1-16, ISBN 978-0-646-50995-2.

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

AUTHOR BIOGRAPHIES

Jim Whittington is a Senior Lecturer in the Department of Electronic Engineering at La Trobe University. His research interests include: development of digital systems and techniques for communication, HF radar, automotive safety and control applications. This particularly includes the real-time hardware implementation of digital signal processing structures in FPGA technologies.

Dr Harvey Ye is a Lecturer in the Department of Electronic Engineering at La Trobe University. His research interests include: digital signal processing, communication systems, and HF radar applications.

Karthik Kamalakkannan is a research associate in the Department of Electronic Engineering at La Trobe University. His research interests include, integrated software/hardware solutions for automotive safety and control, and HF radar applications.

Ngoc Vinh Vu is a PhD student in the Department of Electronic Engineering at La Trobe University. His PhD topic is "In-car Speech Recognition Design and Implementation with Enhancement".

Dr Michael Mason is a Lecturer in the School of Engineering Systems at Queensland University of Technology. His research interests include: speaker identification and verification; speech recognition, enhancement and coding; biometric person authentication; audio coding; and digital signal processing.

Dr Tristan Kleinschmidt recently graduated with a PhD from the School of Engineering Systems at Queensland University of Technology. His PhD topic was “noise-robust automatic speech recognition using speech enhancement”.

Professor Sridha Sridharan is the program leader of the Speech, Audio, Image & Video Technology (SAVIT) research group at Queensland University of Technology. His research interests include: speech enhancement, coding and recognition; speaker recognition and verification; audio coding; lip and face recognition; authentication - biometric person authentication; digital signal processing; and digital communication.