Speech Recognizer with improved detection performance in an Automotive Environment

Ashtosh Sapru, Ravi Lakkundi and Nisar Ahmed
Aricent Communications Pvt Ltd.
Audio and Speech Engineering
Bangalore India
ashtosh.sapru@aricent.com

Abstract

In-car speech recognition is a challenging area of research. The area has been studied [1], and more of is concentrated on addressing the issues in Acoustic Echo Cancellation (AEC). In application such as a voice controlled car audio system , voice commands by the driver are corrupted by audio out of loudspeaker. In this paper we propose and implement a robust, low complex voice controller for car audio system involving an efficient AEC and an effective detection module for a simple speech recognizer. The proposed method has been implemented on an embedded ARM 9T platform [2], with a performance of under 50 Mega Cycles for complete system and is tested in real time environment.

1 Introduction

The desire for having speech recognition and control in automotive applications is a well studied area [3]. Most of the speech control revolve around Isolated word recognition, which is an extensive studied subject [4] and complex algorithms proposed fare excellent with a hit ratio near to 100%. Implementing one of these methods on today's embedded devices is still a challenge, as the algorithms with fairly accurate hit ratio are complex and simpler algorithms do not perform satisfactorily.

A typical Car Speech Interface system used for controlling car audio involves the problem of speech recognition. Speech recognition which is performed using a Command Recognizer (CR) module, is the central part of Interface system. The CR receives the voice commands by the driver for controlling the state of car audio such as play, pause and stop. These commands will be mixed with the audio signal of the car loud speaker, causing a corruption in the performance of command recognizer.

The problem of filtering driver's speech can be solved using an Acoustic Echo Canceller (AEC). AEC uses an adaptive filter to cancel the acoustic echo. An adaptive filter models the impulse response between the loudspeakers and the microphone. Implementing an efficient AEC on an low cost embedded platform can be challenging, the complexity of porting becomes even more critical when a complex speech recognition algorithm is also integrated. The solution to this porting problem can be either to have a complex AEC with a simpler speech recognizer or vice versa. It would be always desirable to have least music contamination of speech.

We here propose a Car Speech Interface system to control car audio. The implemented system contains a simple word recognizer, powered by an efficient adaptive filter and a simple yet highly efficient detection module called False State Detector (FSD). The entire system is implemented on ARM9T processor[2] consuming only an impressive 50 Mega Cycles inclusive of an audio decoder in real time.

2 System Description

Figure 1 shows the block diagram of the proposed car Speech Interface system . The system consists of Car Audio Source which plays out music and a Microphone for

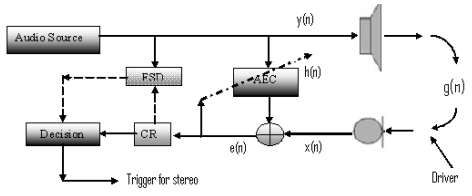

Figure 1. Block Diagram of Speech Interface System

capturing the voice commands of driver. As can be seen from the block diagram there exists an acoustic feedback path between the Car Audio speakers(CA) and the microphone(Mic). As a result the signal getting into Mic is a mixture of signals coming in from CA and the driver(D). The Mic signal x can be written as a summation of echo signal $\hat{y}$ due to impulse response of the car environment and speech v coming from D. If x is sent directly to the Command Recognizer(CR) module, degradation in recognizer performance occurs, due to the influence of $\hat{y}$ which acts as additive noise for the signal x.

Block diagram of Car Speech Interface system shows the use of AEC in removing the undesirable distortion in signal x. The AEC consists of an adaptive filter which models the effect of acoustic feedback path on the car audio signal. The input to the adaptive filter is the signal y and the output of adaptive filter is the echo estimate. The echo estimate $\hat{x}$ is subtracted from the signal x to produce an output e in which the effect of echo signal is marginalized.

The AEC output signal is sent to CR to be analyzed for presence of the spoken command word. We here use an algorithm with low complexity for command word detection based on Dynamic Time Warping(DTW). Mel Frequency Cepstral Coefficients (MFCC) were used as feature vectors because they have been shown to be very effective for speech recognition applications [5]. The feature vectors computed using signal e and reference feature vector are sent as input to DTW module to produce a distance measure which is used for changing the existing state of CA system. When the distance score is less than a threshold a stateflag is set. FSD takes the stateflag as input and produces a decision whether to alter the CA state when voice command of speaker is detected.

2.1 Acoustic Echo Cancellation

AEC operates on signal coming from car audio and signal picked by the microphone. To address the stability issues caused by IIR filters usually a FIR adaptive filter is used in AEC. Normalized Least Mean Square algorithm(NLMS) has been shown to work well for the above problem modelling [6]. However stable adaptation of NLMS filter coefficients remains an issue under conditions of Double Talk (DT). Double talk refers to the situation when Mic signal x consists of acoustic echo $\hat{y}$ and actual driver speech v. A widely used methodology in AEC literature has been the use of a DT detector to stop filter adaptation in DT periods [7]. However, even a small error in establishing the DT periods correctly can cause the adaptive filter to lose convergence. Again if the echo path changes during DT periods, the filter needs to adapt quickly to give satisfactory performance. Recent literature on AEC has seen the development of more robust adaptive algorithms which

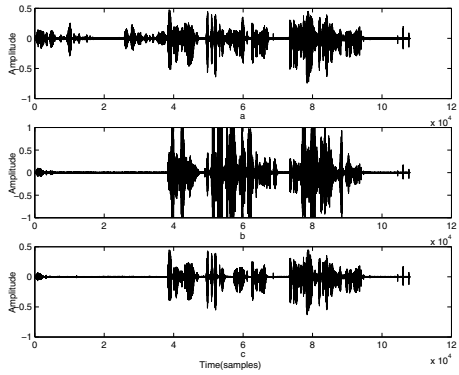

Figure 2. Echo Canceller comparison (a) Microphone input signal, (b) Output using NLMS, (c) Output using Double talk resistant method

work effectively under DT conditions [8]. Due to its double talk resistant feature AEC based upon [8], which works by the minimizing the prediction error was included in the proposed Car Speech Interface system.

Figure 2 shows the behavior of different algorithms under DT conditions as recorded in a moving car. In the plot DT period starts around sample 40000. The algorithm based upon NLMS without any DT protection gives a corrupted output as shown in Figure 2b. From the Figure 2c it can be observed that the performance of DT resistant algorithm is better in the DT period.

2.2 Command Recognizer

The CR analyzes the AEC output signal for the spoken command word. The signal analysis is done by extracting MFCC's as feature vectors. The computed feature vector is matched against already generated and stored reference vector space. The reference vector space is formed by choice of actual command words to be spoken by user. The reference and computed feature vectors are sent to a Dynamic Time Warping module to give a distance score. This score is compared against a threshold ($\mathbf{T}_x$) to generate a state flag which is further analyzed to check for false alarms using FSD.

2.3 False State Detector

It is probable in an acoustic environment that due to a mismatch between the actual echo path and the adaptive modelling of the same,a component of residual echo might be present in the output. This mismatch can occur due to sudden change in echo path. If residual echo is present then there is a possibility that it might be inferred by the Command Recognizer as a spoken command. This situation can

occur due to poor performance of the low complexity CR or when the car audio signal y itself has the command word to be detected. To avoid such false triggers we propose a False State Detector(FSD). FSD works by exploiting the correlation between the residual echo and the car audio sound.

The output $e(n)$ can be written as,

$$\begin{aligned} e(n) &= x(n) - y(n) * h(n), && (1)\\ &= v(n) + y(n) * g(n) - y(n) * h(n), && (2)\\ &= v(n) + \sum_{k=0}^{N-1} y(k)g(k) - \sum_{k=0}^{N-1} y(n)h(n). && (3) \end{aligned}$$

When double talk is not present i.e. $v(n) = 0$, the above reduces to,

$$= \sum_{k=0}^{N-1} y(k)(g(k) - h(k)). \quad (4)$$

where h is the estimated echo path impulse response and g is the actual echo path impulse response.
The cross correlation α_{ye} between y and e is given by,

$$\alpha_{ye} = E[y(n)e(n)] \quad (5)$$

Using eqn. 4 and linearity of expectation operator, the expression for α_{ye} can be reduced as follows,

$$= \sum_{k=0}^{N-1} (E[y(n)y(k)](g(k) - h(k))) \quad (6)$$

Assuming y to be an uncorrelated random process, we can write α_{ye} as,

$$\alpha_{ye} = \sigma_y^2 \sum_{k=0}^{N-1} (g(k) - h(k)) \quad (7)$$

From the above analysis it is clear that when the adaptive filter has not converged, there is significant correlation between the residual and the car audio signal.

The common information between signal y and signal e has been used in designing the False State Detector. The FSD is called when the state flag is set by a particular AEC output frame. To reduce the probability of this being a false alarm when the adaptive filter is converging, we pass the corresponding car audio frame through the command recognizer. If the command recognizer returns a value less than a fixed threshold ($\mathbf{T}_y$), the existing state of CA system remains unchanged.

3 Experiment Results

To quantify the performance of proposed Car Speech Interface system, experiments were performed in a car cabin.

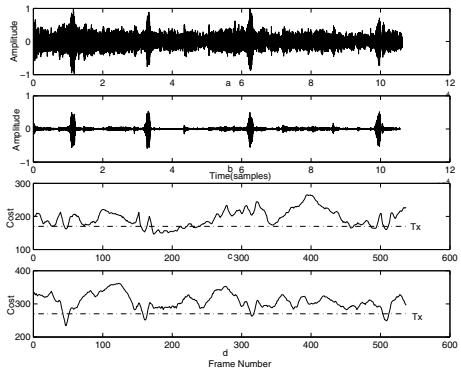

Figure 3. Command Recognizer Output(a) Mic in signal, (b) AEC ouput signal, (c) Unresolved plot using MicIn signal directly, (d)Resolved plot using AEC output

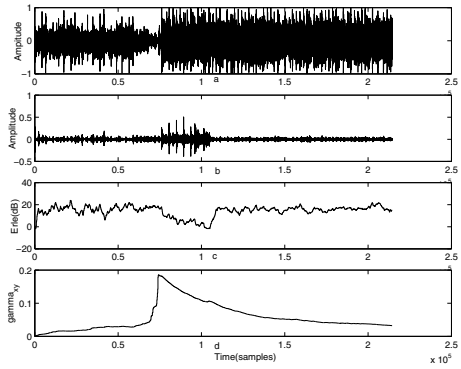

Figure 4. FSD operation (a) Car audio signal, (b) AEC output signal, (c) ERLE vs Time, (d) γ vs Time

The hardware chosen for conducting the experiment was startup kit of OMAP having ARMv4 as the core.The distance between driver and Microphone was about 0.35 meter. The stereo setup was used with a distance of about 1.8 meter between stereo centroid and Microphone.

Figures 3a and 3b show the plot of Mic input and AEC output respectively. Figure 3c shows the plot of distance measure when Mic signal x is sent as input to command recognizer. The plot highlights the difficulty of finding individual command words due to unresolved valleys for this input. Figure 3d shows the plot of distance measure when AEC output is fed to the command recognizer. The plot clearly shows points having marked valleys. The distance measure is compared against ($\mathbf{T}_x$) to give points corresponding to spoken command words.

Figures 4a and 4b show the car audio signal and error

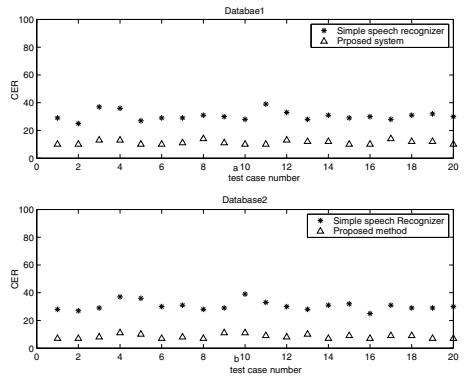

Figure 5. Performance on Database

Table 1. Performance on device

Module	Average Mega Cycles
MP3Decoder	23
AEC	21
CR	5
FSD	1

signal respectively. In Figure 4c we have plotted Echo Return Loss Enhancement(ERLE) as a function of time for the above signals. Figure 4d shows the plot of normalized correlation $\gamma = \alpha_{ye}/\sigma_y^2$ between the car audio signal and error signal. There was no Driver speech while performing the experiment. The sudden drop in ERLE occurs due to a change in echo path impulse response around sample position 78000. Since the correlation curve also rises around same time confirming our supposition that signal in Figure 4a can be used in FSD to detect false alarms.

Figure 5 shows the plot of Command Error Rate(CER) against test case number. Test cases were selected from an in-house database of Hindi and English songs. From the plot it can be seen that there is a significant improvement in CER with a best case improvement from 39.33 $\%$ to 9.89$\%$.

The performance figures of the proposed Car Speech Interface have been highlighted in Table 1. The complete system was implemented and ported on to ARM 9T based hardware. The system was targeted for a low cost processor which set the constraints on the available Mega Cycles. As can be seen from the Average Mega Cycles, AEC is relatively complex. AEC being critical for the application we used a low complex command recognition algorithm. The table also highlights the performance of FSD consuming considerably less mega cycles.

4 Conclusions

In the car environment the voice commands sent to Car Speech Interface system are corrupted by presence of car audio signal. The system works by using an acoustic echo canceller to cancel the acoustic echo captured by the car Microphone. For good command word detection a double talk robust AEC was used. A False state detector was proposed to avoid false alarm cases where the residual music may sometime cause the system to trigger. The proposed FSD helps in checking false alarms due to residual music component by exploiting the fact that correlation between music residual and car audio signal is significant. The proposed system showed a best case CER improvement of around 29$\%$. A future area of study will be to analyze perturbation of reference feature vectors under test conditions.

References

[1] Matassoni M, Omologo M, Zieger C, "Experiments of in-car audio compensation for hands-free speech recognition", *Automatic Speech Recognition and Understanding*, 2003. 2003 IEEE Workshop on Volume , Issue, 30 Nov.-3 Dec. 2003 Page(s): 369 - 374

[2] ARM V4/5 Architecture Reference Manual, http://www.arm.com

[3] http://www.nuance.com/ads/automotive/survey/

[4] Birkenes O, Matsui T, Tanabe K, "Isolated-Word Recognition with Penalized Logistic Regression Machines", *Acoustics, Speech and Signal Processing*. ICASSP 2006 Proceedings Volume 1, Issue , 14-19 May 2006 Page(s):I - I.

[5] Steven B Davis, Paul Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. ASSP-28, No. 4, August 1980

[6] S.Haykin, *Adaptive FIlter Theory*, Prentice-Hall,1996

[7] Jacob Benesty, Dennis R. Morgan, "A New Class of Doubletalk Detectors Based on Cross-Correlation", *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 2, March 2000

[8] Toon van Waterschoot,Geert Rombouts, Piet Verhoeve,Marc Moonen, "Double-Talk-Robust Prediction Error Identification Algorithms for Acoustic Echo Cancellation", *IEEE Transactions on Signal Processing*, Vol. 55, No. 3, March 2007