A Commercial Car Navigation System using Korean Large Vocabulary Automatic Speech Recognizer

Sung Joo Lee, Hoon Chung, Jeon Gue Park, Ho-Young Jung and Yunkeun Lee
Electronics and Telecommunications Research Institute, Daejeon 305-350, South Korea
E-mail: lee1862@etri.re.kr Tel/Fax: +82-42-860-5732/4889

***Abstract*— In this paper, a Korean large vocabulary speech recognizer for an embedded car navigation device is introduced. The proposed speech recognizer identifies 450k point-of-interests within a resource-limited device without serious performance degradation under severe car-noise environments. Before launching the speech recognition application on the Korean retail market, a series of speech recognition tests are conducted in various moving vehicles including sport utility vehicles, recreation vehicles and so on. The on-line 10-best evaluation results of 450k point-of-interest name recognition task show 84.2% accuracy under various driving conditions.**

I. INTRODUCTION

If you are not familiar with the user interface for a small embedded device, it is troublesome to push the alphabet buttons on its touchpad in order to get some information. When it should be done at the same time while driving, it could increase driver distraction. This kind of distraction could make a driver be at risk of being in an accident. The progress in speech recognition technology over the last decades makes it possible to control in-car devices by voice. A driver can select a song to play on an MP3 player or make a phone call using voice control technology while driving, just to name a few. In 2008, Nuance study showed that voice recognition increased car safety by alleviating driver's distraction. But the application of speech recognition technology still remains in relatively small vocabulary size(<60k) because of the limited resource of a current in-car device.

In this paper, we introduce a large vocabulary automatic speech recognizer(ASR) which is embedded in a car navigation device and helps a driver to find a route to destination. Despite of recent progress in speech recognition technology, speech recognition performance can be degraded in severe car-noise environments [1]. If a microphone is located some distance away from a user such as a hands-free application, current automatic speech recognition technology still has difficulty in satisfying user's needs for recognition accuracy. Therefore, it is very important that a speech recognizer should be able to cope with car noises in order to commercialize voice-enabled in-car devices. To do so, a speech preprocessor should be able to remove ambient noise components without speech distortion and detect speech boundaries by identifying speech portions. In addition, acoustic models(AMs) based on Hidden Markov Model(HMM) [2] should be effectively adapted to various noisy conditions which are made by a moving car. In addition, an efficient design scheme of a speech recognition decoder is required in order to apply a large vocabulary speech recognition technology into a resource-limited device.

We propose a large vocabulary speech recognizer which is robust especially in car environment. The speech preprocessor in our system deals successfully with in-vehicle noise by the help of three component technologies: a single channel speech enhancement method based on the widespread Wiener filter [6], an end-point detection(EPD) method with the proposed car-noise robust feature, and the proposed simple speech/non-speech discrimination method. In order to improve speech quality of the conventional Wiener filter [6] for a voice recognizer, a power spectral density(PSD) estimator based on the human auditory model [3] and a voice activity detector(VAD) based on global speech absence probability(GSAP) [10] are proposed. And in order to recognize hundreds of thousands of point-of-interest(POI) names in an embedded device without serious increase of computational complexity and memory requirement, a two-stage decoder based on the human speech recognition(HSR) architecture is adopted [4,5]. In order to notify a hostile driving condition against voice recognition to a user, we propose an environment change detector(ECD) which can estimate on-line signal-to-noise ratio(SNR) and play a warning message on low SNR input signal.

This paper is organized as follows. After describing the proposed speech preprocessor and ECD in Section II, the fast and memory-efficient speech recognition decoding scheme is briefly introduced in Section III. In Section IV, the developed voice-enabled navigator is described. The speech recognition accuracy evaluation results for real data in various driving conditions are described in Section V before the conclusions in Section VI.

II. THE PROPOSED SPEECH PREPROCESSOR

The role of a speech preprocessor is very important to preserve speech recognition performance in real-field applications. Residual noise components and speech signal distortion after speech enhancement often make it difficult to recognize user's voice. And speech recognition accuracy is affected by endpoint detection accuracy. If noise portions are misidentified as speech by the EPD, it often leads to speech recognition performance degradation.

In order to suppress in-vehicle noise components while minimizing speech signal distortion, we analyze the frequency characteristics of various car noise patterns. It is observed that

most in-vehicle noise components are distributed in the frequency range below 2kHz and a lot of vehicle-engine noise components are concentrated in the range below 200Hz. Therefore, a simple high-pass filter with cutoff frequency around 200Hz can successfully reduce vehicle-engine noises while preserving speech components and this scheme of the noise suppression is useful to improve the EPD accuracy in severe car-noise condition. Except the vehicle-engine noise, various types of noises are recorded in a moving car. For example, speech signals are often corrupted by the background noise which is caused by outside traffic, coarse road surface, an air-conditioner, blowing air from an open window, driving noise during speedup, other passenger's babble noise, and etc. In order to remove the negative influence of these adverse noises, we adopt a single-channel speech enhancement algorithm based on the Wiener filter method [6]. In our proposed method, a PSD estimator emphasizes useful frequency components for speech recognition [3] and more accurate PSDs of speech can be obtained by exploiting the VAD which is based on a statistical model, such as GSAP [10].

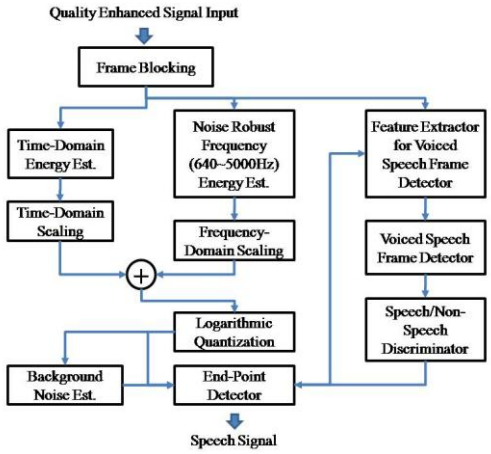


Fig. 1 Bock diagram of the proposed end-point detector

Fig.1 shows the block diagram of the proposed end-point detector with speech/non-speech discriminating function. After investigating the robust frequency range against in-vehicle noise, it is found that the frequency band energy between 640~5000Hz is appropriate for the purpose of speech frame detection in car. The details of the proposed logarithmic time-frequency(TF) energy extraction for the EPD of speech is also described in Fig.1.

Sometimes, burst noise caused by an abrupt driving condition changes makes it difficult to find the boundaries of speech portions among input signal. In the worst case, it leads to the performance degradation of ASR by misidentifying noise portions as speech. Therefore, an EPD should be able to identify substantial speech portions by discriminating between speech and non-speech portions from input signal. In our proposed method, the speech/non-speech discrimination idea is realized by a simple voiced speech frame ratio(VSFR) comparison method for computational efficiency. The VSFR is calculated after separating voiced speech frame from input signal. It is found the following four features for voiced speech frame detection task are efficient and prominent after a series of feasibility study on several features which are corrupted by car noise: Logarithmic TF energy, High-to-low band energy ratio, Zero crossing rate(ZCR) [7], Level crossing rate(LCR) [7].

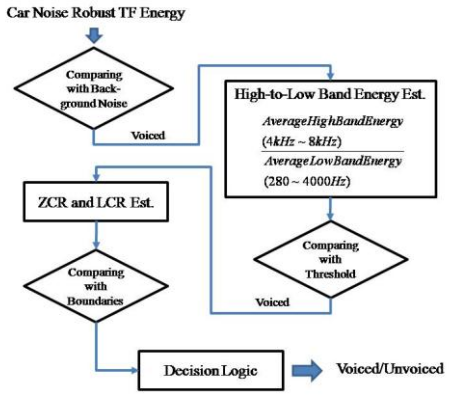


Fig. 2 Bock diagram of the proposed voiced speech frame detector

The high-to-low band energy ratio is obtained according to Fig.2. As shown in Fig. 2, each feature extraction procedure will be started after the current frame is identified as voiced speech on the previous bock to reduce the computational load. The speech/non-speech discrimination procedure in Fig. 1 will be started when a speech starting point is detected by the EPD. The speech/non-speech discriminator will count the number of incoming speech frames after a speech starting point is detected and estimate the corresponding VSFR. If the number of incoming speech frames is enough for a decision making, the discrimination decision will be made by comparing the estimated VSFR with the pre-defined threshold. If the VSFR is less, the burden of incoming speech frames will be rejected and then, the end-point detector will restart searching new speech boundaries. This discrimination process is also applied to the ending boundary of speech portions. Furthermore, the speech/non-speech discriminator should be able to control the end-point detector in order to manage unexpected noise which is caused by abrupt driving condition changes while a user is speaking. In this burst noise condition, the EPD often fails to detect word boundary, especially ending point of speech, because burst noise portions are often misidentified as speech. As a result, this kind of misidentification leads to abnormally long word length and speech recognition performance degradation. To alleviate this problem, when unvoiced frames are detected more than the pre-defined threshold after last voiced speech appearance, the speech/non-speech discriminator stops the EPD process and then, the end-point detector checks whether the incoming frames are speech or not using the corresponding VSFR.

As a result of cooperation of three component technologies: single channel speech enhancement, end-point detection, and speech/non-speech discrimination, the proposed speech

preprocessor can maintain its robustness in various driving conditions.

A novice user who is not familiar with ASR technology sometimes does not understand that ASR accuracy can be affected by severe car-noise. So, in order to prevent this misunderstanding, an ASR system needs to notify a hostile car-noise condition to the user. To do so, we develop an ECD which estimates on-line SNR and plays a warning message of low SNR. The proposed EPD works as follows.

1. Extraction of filter-bank energies from the quality enhanced speech signal during feature extraction process
2. Finding the change points between speech and background noise portions using the Bayesian information criterion(BIC)[8]
3. Estimation of SNR using speech and noise segmentation information
4. Comparison of the SNR and the pre-defined threshold
5. If the SNR is lower than the threshold, play a warning message to notify the user of bad situation for automatic speech recognition

It is supposed that the proposed ECD helps a novice user to understand the technical limit of current speech recognition performance by announcing bad situations for automatic speech recognition.

It is hard to measure the computational complexity of each components of the speech preprocessor including a MFCC feature extractor, because some component modules share the results of operations, such as fast Fourier transform(FFT) power. From the point of view of major computational operations, the proposed speech preprocessor including a MFCC feature extractor requires two real-FFT operations.

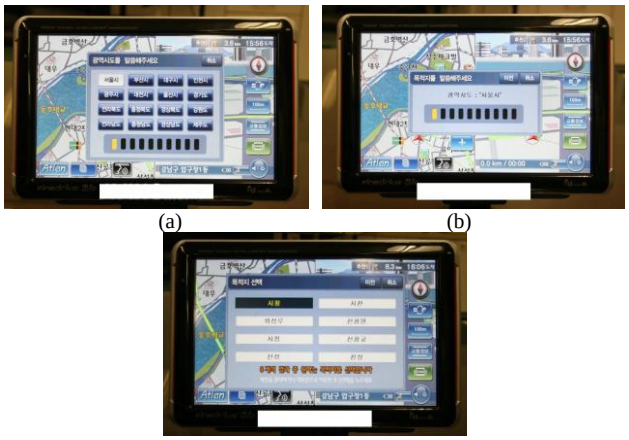
III. Speech Recognition Decoder

Currently, there is an increasing need to access huge amount of databases by voice on embedded devices, e.g., searching a destination POI in a car navigation system or sending a short message on a mobile phone. It is a challenging task to develop a very large vocabulary speech recognition system which works on a resource-limited device since a much faster and memory efficient decoding algorithm must be developed to compensate for hardware limitations.

In this paper, in order to recognize hundreds of thousands of entries in a resource-limited device without serious accuracy degradation of N-best results, we use the fast search scheme based on multi-stage decoding using subspace distribution clustering hidden Markov(SDCHMM)-based AMs. Memory efficiency can be achieved by using the SDCHMM-based AMs. The decoding algorithm is composed of two stages. The first stage is to select rapidly a small set of candidates that are assumed to contain a correct hypothesis with high probability and the second stage re-ranks the candidates by performing acoustic rescoring. Principally, this decoding scheme shares the human speech recognition(HSR) architecture. Multi-layered framework in HSR is completed through a three-stage decoding procedure: acoustic feature to phoneme conversion, phoneme to word conversion and word level rescoring[5]. Computational efficiency is the main advantage of the two stage decoding algorithm which mimics the HSR architecture. The detailed algorithm can be found in Ref. [4]. The reason that the coarse match can achieve significant speed improvement is due to total search space reduction by separating the conventional state-based search space into two independent search spaces, acoustic and lexical spaces. This decoding algorithm improves recognition speed minimum 10-fold compared to the conventional speech recognition systems at a similar level of recognition [4].

IV. Overview Of The Voice-Enabled Navigation System

We developed a commercial voice-enabled car navigator with Korean venture company FINEDIGITAL. If a user wants to find a route to destination using this voice-enabled car navigation system, the speech recognition mode should be activated at first. After the speech recognition mode is on, the navigator plays a beep sound and displays city/province names as shown in Fig.3 (a). After recognizing the city/province name, the navigator displays the screen in Fig.3 (b) and plays a beep sound when it gets ready to recognize point-of-interest(POI) names. Now, the user is able to get route information simply by talking a POI name to the navigator. When 8-best results appear on the screen as shown in Fig.3 (c), the user can select a POI name bar. If the POI name to destination is not shown on the screen, the user can try again by pushing the button on a remote controller.



(a) (b) (c)

Fig. 3 (a) Screen shot after enabling speech recognition mode, (b) Screen shot after recognizing city/province name, (c) 8-best recognition result

V. Experimental Result

In order to train AMs, a huge number of phonetically optimized utterances are recorded from about 1,500 persons. The speech database(DB) is digitalized in 16bit PCM at the sampling rate of 16 kHz and collected in quiet office environment. This training DB is far from the target(various vehicles and driving conditions). Therefore, an AM adaptation procedure is necessary to obtain the matched AMs to in-vehicle environments. 8000 utterances are collected from 80 persons in moving cars(60km/h~100km/h) and the

clean AMs are adjusted by the discriminative AM adaptation method [9]. Mel-frequency Cepstral Coefficients(MFCCs, 13), first and second derivatives are extracted as a speech recognition feature.

To evaluate the speech recognition performance before launching the voice-enabled navigation system on the Korean retail market, a series of speech recognition tests were conducted under various speed and road surface conditions. 30 persons(15 males and 15 females) in various age groups(from 20s to 50s) were engaged. 2801 and 1402 utterances were collected for city/province and for POI recognition test respectively. Various vehicle types such as D and E segment sedans and sports utility vehicles(SUV) were served for the DB collection. We evaluated 1-best recognition performance in the city/province recognition task and 10-best recognition performance in the POI recognition task. This speech DB acquisition is done by the proposed EPD software and the success rate is 98.3% allowing 300ms margin at the start/end point of speech.

The vocabulary size for the city/province recognition task was 16 and the vocabulary size was 450k for the POI recognition task. The evaluation results including EPD errors showed 92.3% accuracy in the city/province recognition test and 84.2% accuracy in the POI recognition test.

TABLE I
EVALUATION RESULTS

Types of Car Engine	16 city/province recognition task	450k POI recognition task
Gasoline Engine	93.2%	83.0%
Diesel Engine	92.9%	86.8%
Total	92.3%	84.2%

Table I indicates the word recognition rate under two different car-engine types. As shown in Table I, the proposed large vocabulary speech recognizer maintains its robustness irrespective of car-engine types.

In order to assess the performance of the ECD module, SNR estimation accuracy test is conducted. For this test, 122 utterances are collected by 2 speakers(1 male, 1 female) under various driving conditions including asphalt and concrete pavement at high speed, over 100km/h.

TABLE II
SNR ACCURACY EVALUATION RESULTS

	SNR(Manual)	SNR(ECD)
Average	6.07dB	6.44dB
Average Diff.	0dB	1.89dB

Table II indicates average SNR values measured manually and automatically, and average of the absolute difference values between them. It is shown that the proposed ECD well estimates SNR from noisy signal.

VI. CONCLUSION

To cope with in-vehicle noise, three core technologies are adopted to the speech preprocessor for automatic speech recognition: single channel speech enhancement, noise robust endpoint detection, and speech/non-speech discrimination. The proposed speech enhancer with the car engine noise removal filter, the PSD estimator based on the human auditory model and the VAD based on statistical model, can suppress in-vehicle noise under various vehicle types, speed, and road conditions. The proposed endpoint detector specialized for car environment can find utterance boundaries even in severe car-noise conditions by cooperating with the proposed speech/non-speech discriminator. In order to employ large vocabulary automatic speech recognition technology into the car navigation device equipped with 600MHz ARM11 core CPU, the fast and memory efficient decoding algorithm based on the HSR architecture is adopted. And the proposed ECD algorithm helps novice user to understand the technical limit of state-of-art speech recognition technologies by notifying bad situations against voice recognition. By integrating all the technologies: speech preprocessor, ECD, and ASR decoder, we realize large vocabulary speech recognizer within a car navigation system. The real-time factor of the proposed voice recognizer is 1.06xRT and total 16Mbyte memory is required under the vocabulary size, 450k. In the future, a robust speech preprocessor using prior knowledge of both of speech and noise under severe car-noise conditions and more efficient speech recognition decoding algorithm covering more than one million vocabulary sizes will be developed.

ACKNOWLEDGMENT

This work is supported by IT R&D program of MKE/IITA [2008-S-001-01, Development of large vocabulary/ interactive distributed/embedded VUI for new growth engine industries].

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, Vol. 16, pp. 261-291, April 1995.
[2] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, Vol. 3, Issue 1, Part 1, pp. 4-16, Jan. 1986.
[3] Rec. ITU-R BS.1387, "Method for Objective Measurements of Perceived Audio Quality," 1998.
[4] H. Chung and I. Chung, "Memory Efficient and Fast Speech Recognition System for Low Resource Mobile Devices," *IEEE Transactions on Consumer Electronics*, Vol. 52, Issue 3, pp. 792-796, Aug. 2006.
[5] O. Scharenborg, "Parallels between HSR and ASR: How ASR can contribute to HSR," *Proc. INTERSPEECH 2005*, pp. 1237-1240, Sept. 2005.
[6] D. Macho, L. Mauuary, B. Noe, Y. Cheng, D. Ealey, E. Jouvet, H. Kellerher, D. Pearce, and F. Saadoum, "Evaluation of a Noise-Robust DSR Front-End on Aurora Database," *Proc. ICSLP 2002*, pp. 17-20, Sept. 2002.
[7] L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," *IEEE Transactions on Acoustics*, Speech and Signal Processing, Vol. 25, Issue 1, pp. 24-33, Feb. 1977.
[8] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, Vol. 32, pp. 111-126, Sept. 2000.
[9] B. Kang, H. Jung and Y. Lee, "Discriminative Noise Adaptive Training Approach for an Environment Migration," *Proc. INTERSPEECH 2007*, pp. 2085-2089, Aug. 2007.
[10] N. S. Kim and JH Chang, "Spectral Enhancement Based on Global Soft Decision", IEEE Signal Processing Letters, Vol. 7, No. 5, pp. 108-110, May 2000.