International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:1, No:7, 2007

Voice Driven Applications in Non-stationary and Chaotic Environment

C. Kwan, X. Li, D. Lao, Y. Deng, Z. Ren, B. Raj, R. Singh, and R. Stern

***Abstract*—** Automated operations based on voice commands will become more and more important in many applications, including robotics, maintenance operations, etc. However, voice command recognition rates drop quite a lot under non-stationary and chaotic noise environments. In this paper, we tried to significantly improve the speech recognition rates under non-stationary noise environments. First, 298 Navy acronyms have been selected for automatic speech recognition. Data sets were collected under 4 types of noisy environments: factory, buccaneer jet, babble noise in a canteen, and destroyer. Within each noisy environment, 4 levels (5 dB, 15 dB, 25 dB, and clean) of Signal-to-Noise Ratio (SNR) were introduced to corrupt the speech. Second, a new algorithm to estimate speech or no speech regions has been developed, implemented, and evaluated. Third, extensive simulations were carried out. It was found that the combination of the new algorithm, the proper selection of language model and a customized training of the speech recognizer based on clean speech yielded very high recognition rates, which are between 80% and 90% for the four different noisy conditions. Fourth, extensive comparative studies have also been carried out.

***Keywords*—**non-stationary; speech recognition; voice commands

I. Introduction

EXISTING speech recognition software such as IBM via Voice or Dragon Naturally Speaking works well in quiet and stationary background noise environments. However, the recognition performance drops quite significantly in crowded and noisy control room, battle stations, emergency room, factory floor, etc. The main reason is that the noise is non-stationary and chaotic. When speech is corrupted by environmental noise, the distribution of the feature vectors of the corrupted speech is no longer similar to the distributions learned from the training data. This mismatch results in mis-classification and poor recognition [1], [2].

To reduce the effect of mismatch, various techniques have been proposed in the literature, which can be broadly categorized as:

- Noise estimation and filtering that reconditions the speech signal or reconstruct speech feature based on noise characteristics [3]-[6];
- On-line model adaptation to reduce the effect of mismatch in training and test environments [7];
- Extraction of speech features robust to noise [8], [9], including features based on human auditory and perception modeling [10]-[12].

To improve the speech recognition rate in chaotic and non-stationary environment, a promising approach has been proposed in the dissertation of Ramakrishnan [1] and master thesis of Selzer [2] at Carnegie Mellon University (CMU). We call this approach Robust Speech Recognition (RSR) method, which consists of two steps. First, the noisy regions of the speech spectrograms (time-frequency plot of speech signals) are identified and deleted. That is, spectral bands with very low signal-to-noise ratios (SNR) in the spectrogram are deleted. Second, the deleted regions are reconstructed, cepstral features are computed, and then speech is recognized. This spectrogram reconstruction part is done by using statistics and speech characteristics in the remaining high SNR bands. Simulations and experiments performed by researchers at CMU [1], [2] demonstrated that RSR yielded the best performance as compared to other techniques in the literature. However, one limiting factor is in step 1. If the noisy regions are wrongly identified and deleted, the speech recognition performance will be degraded.

In [2], a classifier based approach was developed to estimate the unreliable regions in the spectrogram. This process is termed as the identification of spectrographic mask. Extensive simulations in [2] clearly demonstrated the advantages and power of the new method. However, there is still room for improvement.

In this work, we propose a novel system to improve the speech recognition performance in chaotic and non-stationary environment. The core technology will be the RSR method described earlier. However, we will make one important improvement. The key idea is to use a new sensor called General Electromagnetic Movement Sensor (GEMS), which can be attached to the neck, to identify voiced and un-voiced regions in the speech. GEMS was designed and built by Aliph in San Francisco. We purchased one GEMS and used it for an Army project on multi-modal speech enhancement project. Based on the GEMS outputs, we can delete the un-voiced and unreliable spectral bands that contain only the background noise.

The paper is organized as follows. Section II describes our proposed algorithm for enhancing the speech recognition rates

Manuscript received Jun 12, 2005. This work was supported by the U.S. Navy under the Grant N00014-04-M-0324

C. Kwan, X. Li, D. Lao, Y. Deng, and Z. Ren are with Intelligent Automation, Inc., Rockville, MD 20855 USA (X. Li is the corresponding author. phone: 301-294-5200; fax: 301-294-5201; e-mail: xli@i-a-i.com).

B. Raj is with Mitsubishi Electric Research Labs, Cambridge, MA 02139 USA.

R. Singh is with Haikya Corp., Waterwon, MA 02472 USA

R. Stern are with Carnegie Mellon University, Pittsburg, PA 15213 USA.

in non-stationary noise environments. In Section III, we will summarize the experimental results and the comparative studies. Finally, conclusions will be drawn in Section IV.

II. NOVEL SPEECH RECOGNITION APPROACH

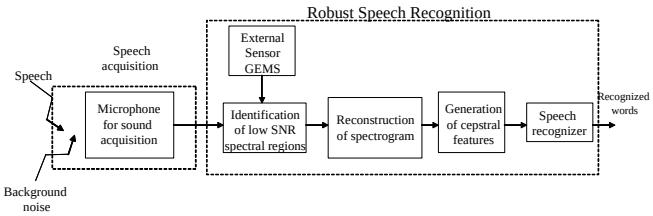


Fig. 1 Architecture of the proposed speech recognition system

As shown in Fig. 1, there are five parts in the proposed robust speech recognition system: an algorithm to identify low SNR regions in the spectrogram, an external sensor called GEMS to assist the identification of low SNR regions that will be deleted, an algorithm to reconstruct spectrogram, a cepstral feature generator, and a speech recognition system.

The main objective of the identification module is to identify low SNR spectral bands in the spectrogram and eliminate them. The background noise level is very high in some applications such as construction sites, helicopter and aircraft cockpits, tanks, factory floor, etc. The presence of noise seriously affects the intelligibility of speech. This is the most critical component of the RSR. Here we propose to exploit external sensors such as General Electromagnetic Movement Sensor (GEMS). The external sensor can provide independent information about where the speech is and this information consequently will help us capture the noise characteristics when there is no speech.

In the spectrogram reconstruction module, the main objective is to optimally reconstruct the regions in the spectrogram. This module has been well developed by Prof. Rich Stern and his students at Carnegie Mellon University. The algorithm known as cluster-based reconstruction has the following advantages. First, it is computationally simple as compared to other techniques. Second, it allows us to generate cepstral features, which have been proven to yield better recognition performance. Third, it yielded the best performance than conventional methods.

In the cepstral feature generation module, a standard approach is used. In the past two years, we used cepstral features in two projects. One is for speaker verification and the other one is for bird classification. The cepstral features yielded excellent recognition performance.

In the speech recognition module, we used the CMU SPHINX speech recognition software, which is open-source. We believe this is more flexible as we can directly adjust some key parameters in the software.

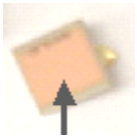
Details of each module will be described below.

A. GEMS for Estimating Low SNR Regions

One problem in identifying the low SNR regions is that it is hard to estimate which portion of microphone signal is speech and which part is not. This problem is even acute in chaotic and non-stationary noise environment. If we can correctly locate speech regions, then it is easy to decide which regions in the spectrogram that we want to delete.

The key idea here is to use an external sensor which can correctly identify speech regions and is independent of background noise. The GEMS developed by Aliph satisfies our needs. We purchased this sensor about 2 years ago and has used it for a multi-modal speech enhancement project for Army Research Laboratory. The GEMS can detect vibratory motion of human tissue. The sensor is an extremely sensitive phase-modulated quadrature motion detector that accurately determines the motion vs. time of one or more moving objects in its field of view. In our application, we will restrict GEMS to detect the motions caused by voiced speech in the sub-glottal or cheek/jaw areas, so we can use GEMS to improve the speech detection accuracy in very noisy environment (>100dB) where conventional speech detection algorithms do not perform well.

Fig. 2 shows the GEMS, which can be attached to the neck or throat area.



GEMS sensor
(less than a quarter)

Fig. 2 GEMS

The data shown in Fig. 3 was collected by GEMS sensor when there was no background noise. The data shown in Fig. 4 was collected from GEMS sensor when background noise level was about 110dB. From Fig. 3 and 4, it can be seen that the GEMS sensor did not collect any noise even though the environment was very noisy.

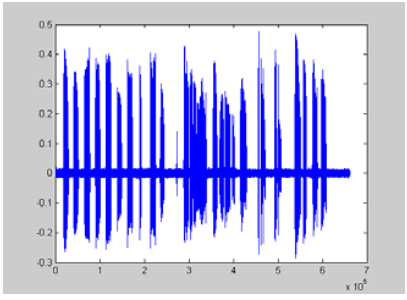


Fig. 3 Speech data collected from GEMS sensor without background noise based on the guidance of GEMS

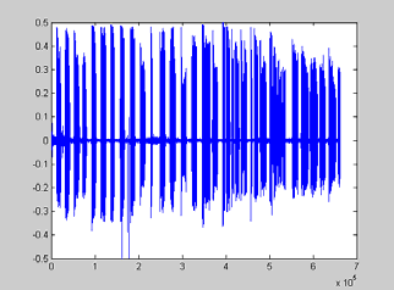
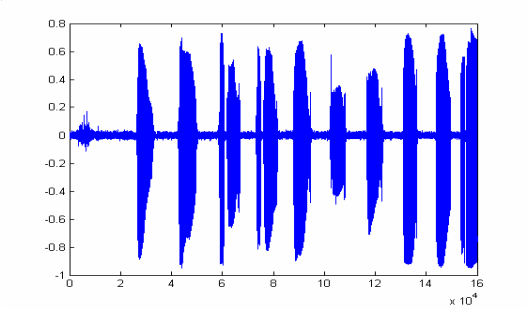


Fig. 4 Speech data collected from GEMS sensor with 110dB background noise with the help of GEMS

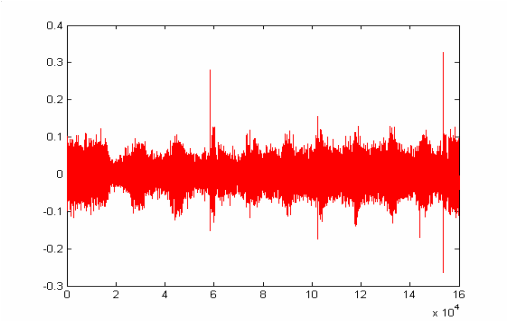
However, the GEMS is also sensitive to placement and attachment. If the location of the sensor is away from the throat area, the signal will be weak. If the attachment is not firm, then wrong indications of speech will occur.

We performed some experiments to investigate the placement and attachment issues. The following bullets summarize the results:

- When a GEMS sensor is placed in the right place, the outputs of the GEMS are strong and clean (Fig. 5)



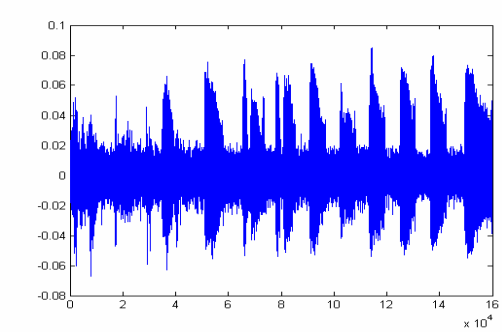
(a) Output of GEMS



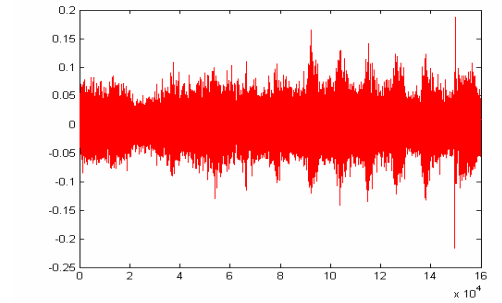
(b) Noisy speech Signal

Fig. 5 Good GEMS outputs when the sensor is properly attached

- When a GEMS sensor is placed in a wrong place of the neck, the output of the GEMS is relatively small and not clean (Fig. 6)



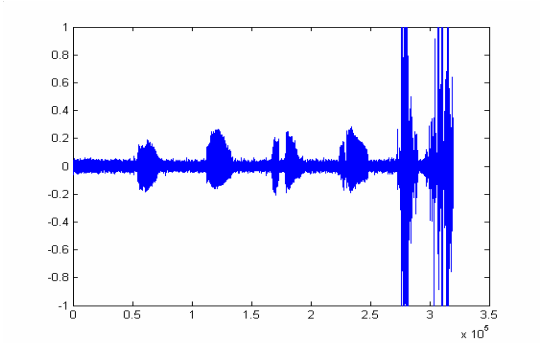
(a) Output of GEMS



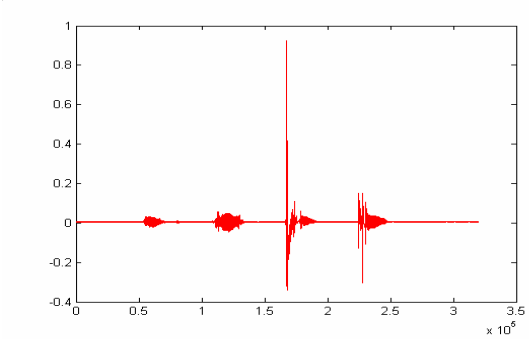
(b) Noisy speech Signal

Fig. 6 Poor GEMS outputs when the sensor was not properly attached to the neck

- From the outputs (Fig. 7 (a)) of the GEMS and outputs of the speech (Fig. 7 (b)), we can see there are two big pulses in GEMS outputs which were caused by head movements and not caused by speech signals.



(a) Output of GEMS



(b) Speech signal

Fig. 7 Head movements may also introduce errors to the GEMS signals, if the GEMS is not properly attached

B. GRATZ: New Mask Estimation with GEMS Data

A new algorithm called GRATZ (GEMS based Multivariable Gaussian Cepstral Normalization) was implemented to use GEMS data to help estimate spectrographic masks. Details of the algorithm will be described in a companion paper [13]. Here we briefly summarize the key idea.

Fig. *8* summarizes the relations between mask estimation and other components of the overall speech recognition system. In the training part, a joint distribution of speech and GEMS features (log-spectral) will be obtained by using speech and GEMS signals collected in the clean environment. In the on-line mask estimation part, both the features from the speech and the GEMS will be used for mask estimation. Spectrogram reconstruction will be done in the log-spectral domain.

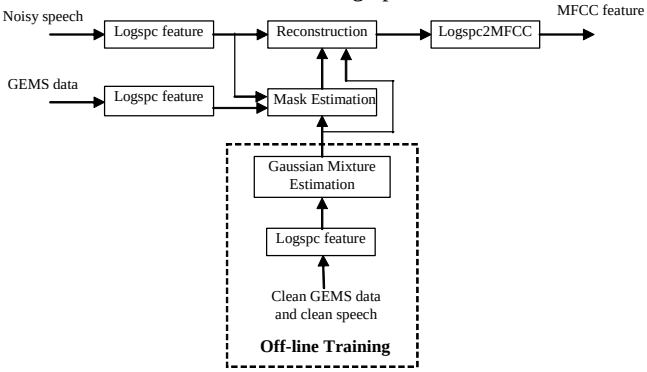


Fig. 8 Relations between mask estimation, feature reconstruction, and speech recognizer

C. Cluster-based Spectrogram Reconstruction (Imputation)

In cluster-based spectrogram reconstruction, the unreliable components of any log-spectral vector are reconstructed based on the reliable components of that vector and the known distribution of the log-spectral vectors of clean speech. This is accomplished by computing a mixture Gaussian distribution from the log-spectral vectors of the spectrograms of a training corpus of clean speech.

The Gaussians of this distribution are all assumed to have diagonal covariances. Once the distribution has been computed, a secondary full covariance matrix is also computed that is common across all the Gaussians in the distribution. The distribution and the covariance matrix can both be computed using the EM algorithm.

In order to reconstruct the missing components of any log-spectral vector $Y(t)$, the unreliable and reliable components of the vector are separated out into two vectors $U(t)$ and $R(t)$. A separate estimate of $U(t)$ is obtained for each of the Gaussians in the mixture based on $R(t)$, the mean of that Gaussian and the global covariance matrix. The estimate is obtained using a bounded MAP procedure. Let us represent the estimate of $U(t)$ obtained for the k^{th} Gaussian as $\hat{U}_k(t)$. We now define the term $P_k(Y(t))$ as:

$$P_k(Y(t)) = \left\{ \prod_r \frac{1}{\sqrt{2\pi\sigma_{k,r}^2}} \exp\left[\frac{-(Y(t,r)-\mu_{k,r})^2}{2\sigma_{k,r}^2} \right] \right\} \left\{ \prod_u \int_{-\infty}^{Y(t,u)} \frac{1}{\sqrt{2\pi\sigma_{k,u}^2}} \exp\left[\frac{-(X-\mu_{k,u})^2}{2\sigma_{k,u}^2} \right] dX \right\} \quad (1)$$

where $\mu_{k,r}$ and $\sigma_{k,r}^2$ represent the mean and variance of the j^{th} dimension in the k^{th} Gaussian. The index r goes over all reliable components of $Y(t)$ and u goes over all unreliable components. We define $P(k|Y(t))$ as

$$P(k \mid Y(t)) = \frac{P_k(Y(t))}{\sum_k P_k(Y(t))} \quad (2)$$

The estimate of the unreliable components of $Y(t)$ is now obtained as

$$\hat{U}(t) = \sum_k P(k \mid Y(t)) \hat{U}_k(t) \quad (3)$$

The estimated values of the unreliable elements are now used to reconstruct a complete spectrogram. The reconstructed spectrogram can either be directly used for recognition, or can be used to derive other features such as cepstra that can be used for recognition. Based on our experience, the log-spectral features yield lower recognition rates as compared to cepstral features.

D. Cepstral Feature Generation Module

As mentioned earlier, speech recognition performance is excellent if cepstral features are used. Here we briefly describe the cepstral features and present some recent results done by us on speaker verification.

The preprocessing subsystem can be described by Fig. 9. In this project, the first 4 blocks in Fig. 9 are not needed because we directly use reconstructed lo spectral features to generate the cepstral coefficient.

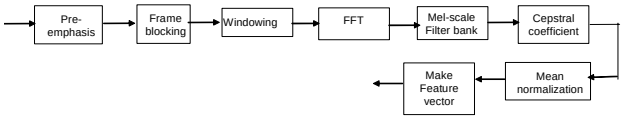


Fig. 9 Cepstral feature generation subsystem

The purpose of feature extraction is to convert each frame of speech into a sequence of feature vectors. In our system, we

use cepstral coefficients derived from a Mel-frequency filter bank to represent a short-term speech spectra. The digital speech data is first preprocessed (pre-emphasized, set to overlapped frames and windowed) and then Mel Frequency Cepstral Coefficient Analysis is applied. Typically feature extraction process compresses around 256 samples of speech data down to between 13 to 40 features.

Mel Frequency Cepstral Coefficients features, $\{c_i\}$, are obtained by taking the Discrete Fourier Transform (DCT) of log MFCCs, $\{A_j\}$, as shown below:

$$c_i = \sum_{j=1}^{N} A_j \cos(\frac{\pi i}{N}(j-0.5)) \quad (4)$$

E. Speech Recognition Module

We have decided to use the CMU open-source SPHINX-3 as the speech recognizer [14]. One advantage is that we can make modification to the recognizer. Another advantage is that, if this research goes to the product stage, the cost of the product will be small.

III. Main Results

A. Data Collection Experiments

We defined the task to be the Navy acronym recognition. The scenario is that a Navy pilot may want to know the meaning of some acronyms and by interacting with a speech recognition system, he can get this information.

The transcripts of acronyms are taken from [19]. The subset we used composed of total 298 sentences with vocabulary size of 521. A few example sentences are listed below:

AAM : Air-to-Air Missile
JPATS: Joint Primary Aircraft Training System
In Commission: Vessel is in active service, operational, with crew assigned

The noise data is available at [20]. Four types of noise were taken from Noisex-92 database:

Cockpit noise (buccaneer jet traveling at 190 knots
Destroyer operations room background noise
Factory noise (plate-cutting and electrical welding equipment)
Speech babble noise (100 people speaking in a canteen).

We played each type of noise at different SNR levels (clean, 25 dB, 15 dB and 5 dB) during recording. Stereo speech data, one channel for microphone and one channel for GEMS sensor, were quantized at 16 bits per sample and sampled at 32 kHz. Two native male speakers', total 7.5 hours, speech were recorded.

B. Speech Recognition Performance of Different Methods

1) Acoustic Model (AM) Training

The CMU Sphinx-3 continuous density Hidden Markov Model (HMM) system was used as the speech recognizer. HMMs with 5000 tied states, each modeled by a mixture of 8 Gaussians, were trained by using Resource Management data, which contains more than 25,000 utterances, spoken by more than 160 speakers.

2) Language Model Training

Language model (LM) is used for speech recognition decoding. It is well known that a good language model is a crucial part of modern speech recognizers. Effective LMs can result in improvements in recognition accuracy of different algorithms.

The task of language model is to estimate the probability of a work sequence $W = \{w_1, w_2, ..., w_n\}$. The uni-gram model accorded equal probability to all the words in the recognition vocabulary. Tri-gram assumes the probability of current word depends only on previous two words.

$$P(W) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)\cdots P(w_n \mid w_{n-1} w_{n-2})$$

In order to train the LM, probability masses of N-grams were redistributed by the Turing discounting strategy [15].

3) Recognition Performance of Different Methods by Using a Trigram Language Model

We first used unigram LM in the speech recognizer. The acoustic model was obtained by using the Resource Management corpus, which is general for all speakers. The improvement in recognition rate was not significant.

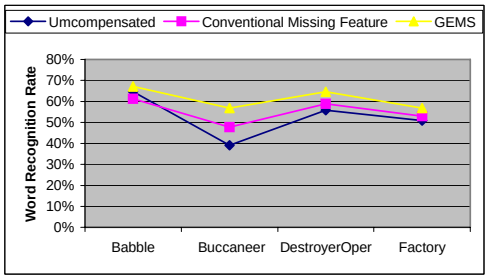


Fig. 10 Speech recognition performance without customized training for the 5 dB case.

Here we will investigate the performance of using a better LM known as trigram, which models the relationship between three consecutive words. Fig. 10 summarizes the results. Three methods were compared. One is uncompensated case. The second one is the conventional approach without GEMS. The third one is the GEMS based spectrogram reconstruction. Some initial results have been report in [16], [17]. It can be seen that, for the 5 dB case, an average of 9% improvement in recognition rate between the case of with GEMS and the uncompensated case has been observed. In the Buccaneer case, we have even observed 17% improvement.

C. GEMS based Speech Recognition Performance with Customized Training

In Section III B (3), we presented some speech recognition results based on GEMS. There, no customized training was used. A general acoustic model based on the Resource Management corpus was used. Here we present some results by allowing some customized training, which was done by using clean speech signals of the speaker.

1) Results of Using 100 Sentences to Train

Here we used 100 sentences of clean speech in our Navy acronym database to train the acoustic model in the speech recognizer.

Fig. 11 summarizes the recognition results by using customized training. Trigram LM was used. It can be seen the average recognition rate has been increased to about 70% by using the GEMS based approach under 5 dB condition. Without compensation, the recognition was quite low (less than 50% in most cases).

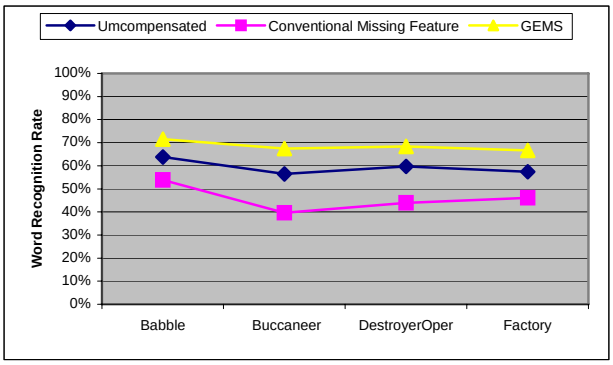


Fig. 11 Speech recognition performance with customized training for the 5 dB case. 100 sentences for training

2) Results of Using All 298 Sentences to Train

Here we investigate the performance of different recognizers when we used all 298 sentences of clean speech to train the acoustic model.

Fig. 12 summarizes the recognition rates of using 298 sentences of clean speech for getting the customized acoustic model. The average recognition rate is over 80% for the GEMS method under 5 dB noise condition.

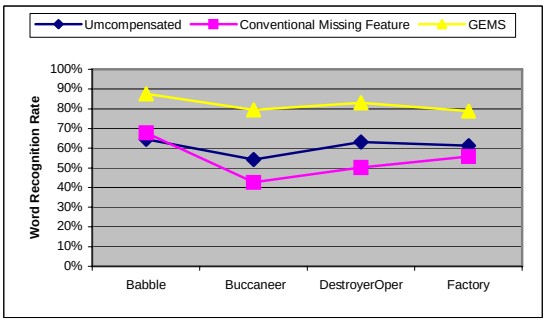


Fig. 12 Speech recognition performance with customized training for the 5 dB case. 298 training sentences

3) Performance Trend of Customized Training

The results shown earlier were obtained using the Sphinx-3 recognizer that was trained by the Resource Management (RM) database. The error rate is quiet high. For example, for SNR=5 dB, and the noise is Destroyer operating room noise, the error rate was 67.42%. This may be due to the fact that the speech characteristics in RM data are very different from those in our database (Navy acronym).

For commercial speech recognition software such as IBM Via Voice and Microsoft Speech Recognizer, training is required to achieve high recognition rate even in quiet office environment. For the challenging noisy environment, we think training is very important as well.

In Section III.C, we have seen two cases that demonstrated good recognition performance by using 100 sentences and 298 sentences for customized training. Here we will produce more results to generate performance trend information by using customized training.

Fig. 13 and Fig. 14 show the word recognition rates for the cases when Sphinx-3 was used to recognize the 151-th to 298-th sentences. Six cases were tested under which the recognizer was trained differently by: RM database, 25 sentences (the first to the 25-th sentences), 50 sentences, 100 sentences, 150 sentences and 298 sentences. Except for the last case, the test sentences are different from the training sentences. It can be seen that with more sentences in the customized training, we can clearly see better recognition rates.

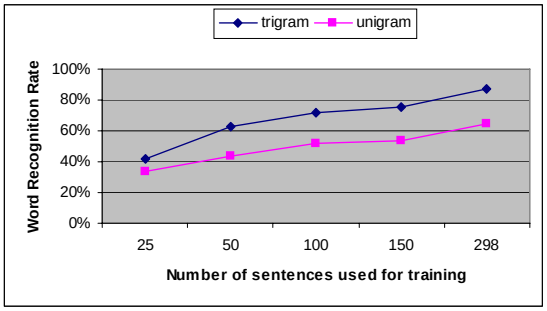


Fig. 13 Performance trend of customized training. More training sentences yielded better recognition rate

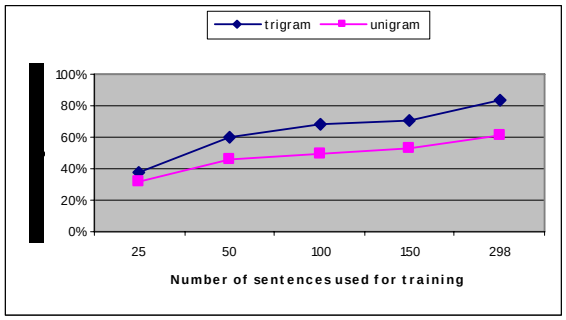


Fig. 14 Performance trend of customized training. More sentences yielded better recognition rate

D. Comparative Studies

Here we present an alternative algorithm from the literature for enhancing speech recognition rate in noisy environment. The main advantage of this approach is that it can work in both stationary and non-stationary environments.

1) RASTA-PLP Algorithm for Robust Feature Extraction

In speech recognition, many different feature representations of the speech signal have been explored. The most popular feature representation currently used is the Mel-frequency Cepstral Coefficients (MFCC). We briefly described MFCC in Section II. Another popular speech feature representation is known as Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP). PLP [11] proposed by Hynek Hermansky is a way of warping spectra to minimize the

differences between speakers while preserving the important speech information. RASTA [18] is a separate technique that applies a band-pass filter to the energy in each frequency subband in order to smooth out the short-term noise variations and to remove any constant offset in the speech channel.

2) Experiments

The Sphinx-3 recognizer was used for the recognition experiments. The recognizer was trained using clean utterances from the recorded Navy acronym database. Only microphone speech from one native speaker was used for both training and testing. Due to the limited amount of data, context independent phone model was trained using a mixture of 8 Gaussians. A trigram model was used in the experiments.

The test set consisted of the data corrupted to SNR of 5 dB by four different kinds of noises: babble, destroyer operations room, Factory, and Buccaneer. As a comparison, a regular conventional MFCC features were used as a baseline. MFCC speech feature includes cepstral and energy term, while as only cepstral is used in RASTA-PLP feature. No features from GEMS sensor signal were used in this experiment.

From Fig. 15, we can see that the recognition rates for all cases have been increased. For example, the recognition rate has been increased from 70% to 80 % in the Babble noise case.

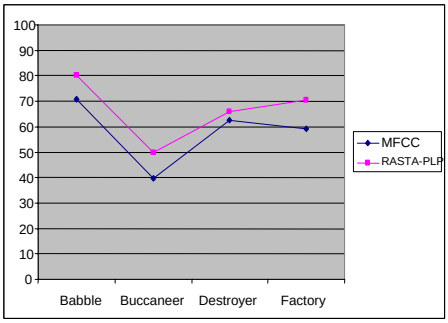


Fig. 15 Word recognition accuracy on Navy acronym Data: MFCC vs. RASTA-PLP with Trigram. 5dB SNR, and 4 noisy conditions (Babble, Buccaneer, Destroyer, and Factory)

IV. Conclusions

In this research, we have clearly demonstrated the proposed approach of improving speech recognition rate under non-stationary and chaotic noise environments. Extensive simulations and comparative studies were performed. Fig. 16 and Fig. 17 summarize some key findings. Specifically, we have observed that:

- The GEMS did improve the speech recognition performance. For example, 16% improvement was achieved for the Buccaneer jet noise case under 5 dB SNR level. The overall recognition rate is about 60% for uncustomized training and 80% for customized training.
- Customized training using clean speech from a specific speaker can improve the recognition rate by about 20%. Now the recognition rate for 5 dB case is over 80%. From Fig. 16 and Fig. 17, we can clearly see the impact of customized training.
- The language model (LM) plays an important role in improving the speech recognition rate. In general, trigram LM improved the recognition rate by about 20% as compared to that of unigram LM.
- RASTA-PLP approach achieved good recognition results. However, it was inferior to the GEMS based method.
- The conventional method is good for stationary noise and is not suitable for non-stationary noise environments.

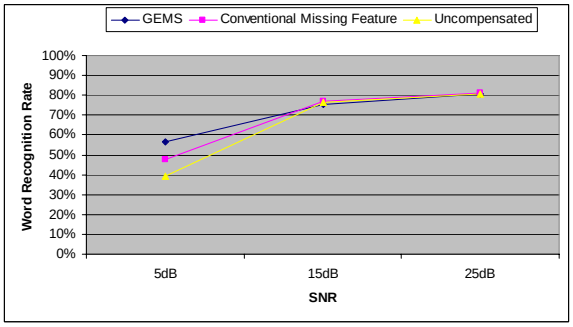


Fig. 16 Speech recognition performance of uncustomized training. Noise environment: Buccaneer. LM: Trigram

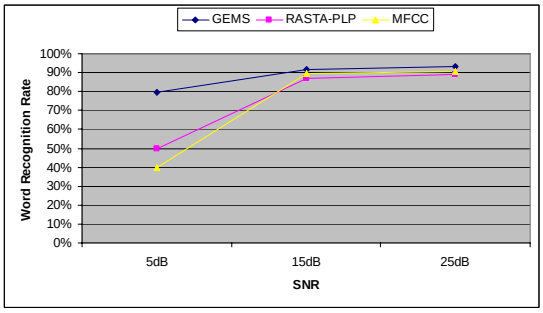


Fig. 17 Speech recognition performance of customized training. Noise environment: Buccaneer. LM: Trigram

Acknowledgment

The authors would like to thank Mr. John Charles (the topic chairman for this program) of the Naval Surface Warfare Center, Bethesda, MD., for his enthusiasm and support throughout this research.

References

[1] B. R. Ramakrishnan, *Recognition of Incomplete Spectrograms for Robust Speech Recognition*, Ph.D. dissertation, Dept. Electrical and Computer Engineering, Carnegie Mellon University, 2000.

[2] M. L. Seltzer, B. Raj, and R. M. Stern, "Classifier-Based Mask Estimation for Missing Feature Methods of Robust Speech Recognition," *Proc. of the International Conference of Spoken Language Processing*, Beijing, China, October, 2000.

[3] S.V., Milner, B.P, "Noise-adaptive hidden Markov models based on Wiener filters", *Proc. European Conf. Speech Technology*, Berlin, Vol. II, pp.1023-1026, 1993.

[4] "Acoustical and Environmental Robustness in Automatic Speech Recognition". A. Acero. Ph. D.Dissertation, ECE Department, CMU, Sept. 1990.

[5] Singh, R., Stern, R.M. and Raj, B., "Signal and Feature Compensation Methods for Robust Speech Recognition," CRC Handbook on Noise Reduction in Speech Applications, Gillian Davis, Ed. CRC Press, 2002.
[6] Singh, R., Raj, B. and Stern, R.M., "Model Compensation and Matched Condition Methods for Robust Speech Recognition," CRC Handbook on Noise Reduction in Speech Applications, Gillian Davis, Ed. CRC Press, 2002.
[7] Nadas, A., Nahamoo, D. and Picheny, M.A, "Speech recognition using noise-adaptive prototypes", *IEEE Trans. Acoust. Speech Signal Process.* Vol.37, No. 10, pp-1495- 1502, 1989.
[8] Mansour, D. and Juang, B.H, "The short-time modified coherence representation and its application for noisy speech recognition", *Proc. IEEE Int.. Conf. Acoust. Speech Signal Process.*, New York, April 1988.
[9] S. Chakrabartty, Y. Deng and G. Cauwenberghs, "Robust Speech Feature Extraction by Growth Transformation in Reproducing Kernel Hilbert Space," *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP'2004)*, Montreal Canada, May 17-21, 2004.
[10] Ghitza, O., "Auditory nerve representation as a basis for speech processing", in *Advances in Speech Signal Processing*, ed. by S. Furui and M.M.Sondhi (Marcel Dekker, New York), Chapter 15, pp.453-485, 1992.
[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoustic Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, Apr. 1990.
[12] Y. Deng, S. Chakrabartty, and G. Cauwenberghs, "Analog Auditory Perception Model for Robust Speech Recognition," *Proc. IEEE Int. Joint Conf. on Neural Network (IJCNN'2004)*, Budapest Hungary, July 2004.
[13] B. Raj et al., GRATZ Algorithm Summary, to be submitted.
[14] "Sphinx-3 s3.3 Decoder", Mosur K. Ravishankar (*aka* Ravi Mosur), Sphinx Speech Group, CMU.
[15] Slava M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 35(3), pp. 400-401, March, 1987.
[16] Bhiksha Raj and Rita Singh, "Feature compensation with secondary sensor measurements for robust speech recognition," *Proc. EUSIPCO 2005*, Antalya, Turkey, August 2005.
[17] Bhiksha Raj, Rita Singh and Paris Smaragdis, "Recognizing speech from simultaneous speakers," *Proc. INTERSPEECH 2005*, Lisbon, Portugal, September 2005.
[18] H . Hermansky, and N. Morgan, "RASTA processing of speech", *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 4, pp. 578-589, Oct. 1994.
[19] http://www.btinternet.com/~a.c.walton/navy/smn-faq/smn2.htm
[20] http://spib.rice.edu/spib/data/signals/noise/

Chiman Kwan received his B.S. degree in electronics with honors from the Chinese University of Hong Kong in 1988 and M.S. and Ph.D. degrees in electrical engineering from the University of Texas at Arlington in 1989 and 1993, respectively. From April 1991 to February 1994, he worked in the Beam Instrumentation Department of the SSC (Superconducting Super Collider Laboratory) in Dallas, Texas, where he was heavily involved in the modeling, simulation and design of modern digital controllers and signal processing algorithms for the beam control and synchronization system. He received an invention award for his work at SSC. Between March 1994 and June 1995, he joined the Automation and Robotics Research Institute in Fort Worth, where he applied neural networks and fuzzy logic to the control of power systems, robots, and motors. Since July 1995, he has been with Intelligent Automation, Inc. in Rockville, Maryland. He has served as Principal Investigator/Program Manager for more than 65 different projects, with total funding exceeding 20 million dollars. Currently, he is the Vice President, leading research and development efforts in signal/image processing, and controls. He has published more than 39 papers in archival journals and has had 90 additional refereed conference papers. He is a senior member of the IEEE.

Xiaokun Li received his BS and MS degree in electrical engineering from Xian Jiaotong University, China, 1992 and 1995 respectively. He obtained his Ph.D. degree in electrical engineering from University of Cincinnati, Ohio, in 2004. From 1995 to 1999, he was an assistant professor at Xian Jiaotong University. He worked as a visiting researcher at SCR (Siemens Corporate Research), Princeton, NJ, in 2002, and MERL (Mitsubishi Electronic Research Labs), Cambridge, MA, in 2003. Since 2004, he has worked with Intelligent Automation Inc as a research engineer. His research interests include signal/image processing and analysis, optical/electronic imaging, medical imaging, computer vision, machine learning, pattern recognition, artificial intelligence, real-time system, and data visualization. He is a member of the IEEE, SPIE, and Sigma Xi.

Debang Lao received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China, China in 1988 and 1993, respectively. He received the Ph.D. degree from New Jersey Institute of Technology, Newark, USA in 2003. Currently he is working on various digital signal processing projects at Intelligent Automation, Inc.

Yunbin Deng received his B.S. degree in Control Engineering from Beijing University of Aeronautics and Astronautics in 1997, his M.S. in Electrical Engineering from Institute of Automation, Chinese Academy of Sciences in 2000, and his M.S.E. in Electrical and Computer Engineering (ECE) form Johns Hopkins University (JHU) in 2002, respectively. He is a Ph.D. candidate at ECE, JHU. He joined Intelligent Automation, Inc. in Rockville, Maryland as a research engineer in 2005. Mr. Deng's research interests include language and speech processing, robust and non-native speech recognition, dialog system, mixed-signal VLSI circuits and system, and machine learning. He won Outstanding Overseas Chinese Student Award by Chinese Scholarship Council at 2003. He is a student member of the IEEE and SPIE. Mr. Deng is a paper reviewer for IEEE Transaction on Audio and Speech Processing, IEEE Transaction on Circuit and System, and Circuit, System, and Signal Processing Journal.

Zhubing Ren received her B.S. (1983) and the M.S. (1989) degrees in Electrical Engineering from the Beijing University of Aeronautics and Astronautics, China. She also obtained a M.S. degree in Electrical Engineering in 2000 from the Johns Hopkins University. From 1983 to 1986 and 1989 to 1995, she worked as a design engineer at Beijing Institute of Radio measurement in China. From 1995-1997, she worked at Glocom Inc, USA, where she was a design engineer and involved in several projects that were related to satellite communication terminals. Since 1997, she has been with Intelligent Automation, Inc (IAI), USA, where she is currently a senior electronic engineer. Ms. Ren's research interests include signal processing, image processing, and fault diagnostics and applications. Over the last 8 years with IAI, she has worked on many different research projects in the above areas funded by various US government agencies such as DoD and NASA. She has also published over 10 journal and conference papers in the related areas.

Bhiksha Raj received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, Pennsylvania, in May 2000. Since 2001, he has been working at Mitsubishi Electric Research Laboratories, Cambridge, Massachusetts. He works mainly on algorithmic aspects of speech recognition, with special emphasis on improving the robustness of speech recognition systems to environmental noise. His latest work was on the use of statistical information encoded by speech recognition systems for various signal processing tasks. He is a member of the IEEE.

Rita Singh received the B.Sc. (Hons.) degree in physics and the M.Sc. degree in exploration geophysics, both from the Banaras Hindu University, India. She received the Ph.D degree in geophysics in 1996 from the National Geophysical Research Institute of the Council of Scientific and Industrial Research, India. She is a Member of the Research Faculty at the School of Computer Science, Carnegie Mellon University (CMU), Pittsburgh, PA, and a Visiting Scientist with the Media Labs, Massachusetts Institute of Technology, Cambridge. From March 1996 to November 1997, she was a Postdoctoral Fellow with the Tata Institute of Fundamental Research, India, where she worked with the Condensed Matter Physics and Computer Systems and Communications Groups. During this period, she worked on nonlinear dynamical systems and signal processing as an extension of her doctoral work on nonlinear geodynamics and chaos. Between November 1997 and February 2004, she has been affiliated with the Robust Speech Recognition and SPHINX Groups at CMU, and has been working on various aspects of speech recognition including core HMM-based recognition technology, automatic learning techniques, and environmental robustness techniques for speech recognition. Since 2004, she became the CEO of Haikya Corp., which focuses on speech recognition products.

Richard M. Stern received the S.B. degree from the Massachusetts Institute of Technology (1970), the M.S. degree from the University of California, Berkeley (1972), and the Ph.D. from MIT (1977), all in electrical engineering.

He has been on the faculty of Carnegie Mellon University since 1977, where he is currently a professor in the Electrical and Computer Engineering, Computer Science, and Biomedical Engineering Departments and the Language Technologies Institute. Much of his current research is in spoken language systems. He is particularly concerned with the development of techniques to make automatic speech recognition more robust with respect to changes in environmental and acoustical ambience. He has also developed sentence parsing and speaker adaptation algorithms for earlier CMU speech systems. In addition to his work in speech recognition, he also maintains an active research program in psychoacoustics, where he is best known for theoretical work in binaural perception. He is a member of the IEEE.