# Feature Compensation for Speech Recognition in Severely Adverse Environments due to Background Noise and Channel Distortion

*Wooil Kim and John H. L. Hansen*

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{wikim,John.Hansen}@utdallas.edu, http://crss.utdallas.edu

## Abstract

This paper proposes an effective feature compensation scheme to address severely adverse environments for robust speech recognition, where background noise and channel distortion are simultaneously involved. An iterative channel estimation method is integrated into the framework of our Parallel Combined Gaussian Mixture Model (PCGMM) based feature compensation algorithm [1]. A new speech corpus is generated which reflects both additive and convolutional noise corruption. The channel distortion effects are obtained from the NTIMIT and CTIMIT corpora. Evaluation of objective speech quality measures including STNR, PESQ, and speech recognition shows that the generated speech corpus represents highly challenging acoustic conditions for speech recognition. Performance evaluation of the proposed system over the obtained speech corpus demonstrates that the proposed feature compensation scheme is significantly effective at improving speech recognition performance with presence of both background noise and channel distortion, comparing to the conventional methods including the ETSI AFE.

**Index Terms**: channel estimation, feature compensation, corpus generation, PCGMM, robust speech recognition.

## 1. Introduction

Mismatch between training and operating conditions of an actual speech recognition system is one of the primary factors severely degrade recognition accuracy. Background noise, microphone mismatch, communication channel, and speaker variability are major sources of such mismatch. Recently, as mobile device such as a smart phone is getting highly popular, speech recognition technology via the mobile system is becoming more challenging, since a range of background noise and time-varying channel effects make the recognition condition more difficult, which can be represented as Fig. 1. This paper focuses on an effective feature compensation scheme for robust speech recognition in a severely adverse environment where additive background noise and channel distortion are simultaneously present.

To minimize the acoustic mismatch, extensive research has been conducted in recent decades, which includes many types of speech/feature enhancement methods such as Spectral Subtraction, Cepstral Mean Normalization (CMN), and variety of feature compensation schemes. Various model adaptation techniques have been successfully employed such as the Maximum A Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR), and Parallel Model Combination (PMC) [1]-[4]. Recently, missing-feature methods have shown promising results.

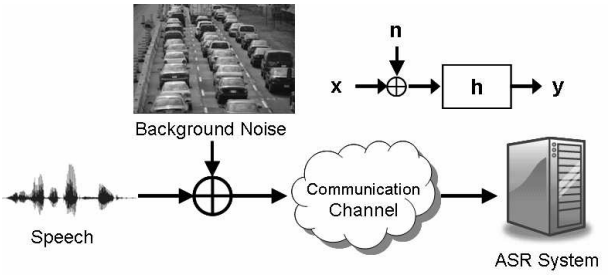


Figure 1: *Speech corruption model with background noise and channel distortion.*

As the real-life condition for speech recognition gets more adverse, a standard speech corpus for developing robust speech recognition algorithm is highly in demand, which simultaneously reflects background noise and channel distortion. Aurora 2.0 database [5] includes a part which contains two types of communication channel effects (i.e., G.712 and MIRS) together with additive background noise, however its channel distortions are not very severe, therefore their effects are dominated by the additive background noise even at a high SNR[1]. The NTIMIT, CTIMIT, and WTIMIT corpora represent highly distorted channel effects which can be observed in real-life communication conditions, but they do not include various types of additive background noise [6]-[8].

In this paper, we develop a speech corpus which simultaneously reflects the additive background noise and channel distortions. The channel effects are estimated from the NTIMIT and CTIMIT corpora, and they are again applied to the TIMIT corpus which is corrupted by background noise. We also propose a feature compensation method to effectively address both background noise and channel distortion for improving speech recognition performance. The proposed system is based on the Parallel Combined Gaussian Mixture Model (PCGMM) method [1], where noise-corrupted speech model is obtained by a model combination method using clean speech and noise models. In this study, an iterative channel estimation method is proposed and integrated into the framework of the PCGMM-based scheme. The proposed system is evaluated over the gen-

This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067 (approved for public release, distribution unlimited), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

[1]Our preliminary experiment shows recognition performance over clean speech only with the channel effects of AURORA 2.0 (without background noise) is comparable to the clean speech condition.

erated speech corpus.

## 2. Corpus Generation

This section presents a procedure of corpus generation employed in this study. The test part of the TIMIT speech corpus is used, which consists of 1680 utterances for 168 different speakers. In order to obtain channel distortion effects, the NTIMIT and CTIMIT corpora are used, where reasonably severe convolutional noise components are included. The speech samples are down-sampled to 8 kHz. To implement the noise corruption model as shown in Fig. 1, first the clean TIMIT speech samples are corrupted by adding background noise samples. In this study, we use car noise and speech babble noise samples which are obtained from AURORA 2.0 database [5]. The location of the noise segment with the same length as the target speech sample is randomly determined from the original noise sample, and then the obtained noise segment is added to the speech sample at 10 dB SNR.

To apply the channel distortion effect to the noise-added speech sample, the channel effect is estimated from the channel-distorted speech sample from the NTIMIT and CTIMIT corpora [7][8]. The target channel-distorted speech sample and its corresponding clean speech sample as a reference need to be aligned in time domain, since the length of speech samples in the NTIMIT and CTIMIT are not identical to the original TIMIT clean speech samples. A simple correlation method was employed in this study. Using the time-aligned channel-distorted and clean speech samples, a transfer function of the channel effect is estimated at every frame. The frame (i.e., window) and overlap sizes are 32 msec (256 samples) and 16 msec respectively. In this study, the transfer function is estimated in the frequency domain. We use a smoothed version of the channel transfer function over the past 15 frames to obtain a more stable function value. To minimize an over-estimate of the channel effect during the non-speech duration, the transfer function is estimated only from the speech segments. The obtained channel function is again applied to the corresponding noise-added speech sample. An average of the obtained transfer functions over the entire speech segments is applied during the non-speech segments of the noise-added speech sample.

Fig. 2 presents examples of (a) clean, (b) corrupted only by additive car noise (TIMIT + car noise), and corrupted speech by both additive and convolutional noise ((c) NTIMIT + car noise and (d) CTIMIT + car noise). The speech samples (c) and (d) were obtained by the corpus generation procedure presented in this section. It can be seen that the speech signals (including background noise segments) are severely degraded by applying the channel effects to the speech corrupted by additive noise in cases of (c) NTIMIT+NOISE and (d) CTIMIT+NOISE. The objective speech quality measures including Signal-to-Noise-Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) are evaluated over the obtained speech corpora, and the details will be discussed in Sec. 4.

## 3. PCGMM-Based Feature Compensation Integrated with Channel Estimation

As presented in Fig. 1, input speech signal is assumed to be characterized in time domain as

$$y(t) = (x(t) + n(t)) * h(t) = x_h(t) + n_h(t). \qquad (1)$$

In this study, an iterative channel estimation method is integrated into our previously proposed PCGMM-based feature compensation scheme to effectively address the background noise $n(t)$ and channel distortion $h(t)$. The following sections present the complete feature compensation algorithm including the proposed channel estimation method. Now all signals represent the feature vectors of Mel-Frequency Cepstral Coefficients (MFCC, c0-c12).

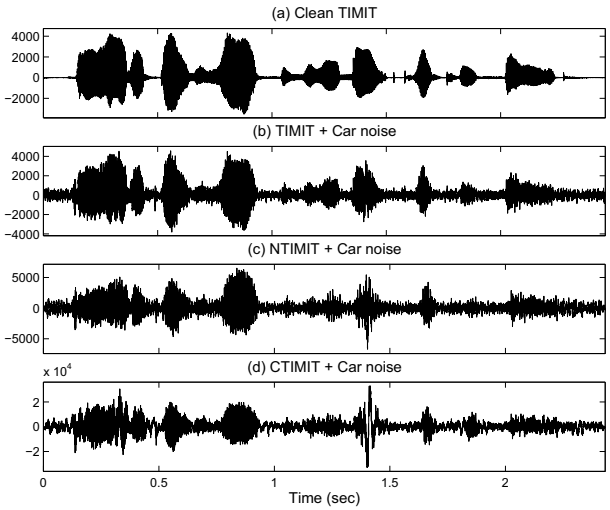


Figure 2: *Example of speech sample obtained by corpus generation: (a) clean condition, (b) corrupted only by car noise, and (c) & (d) corrupted both by car noise and channel distortion.*

### 3.1. Step 1: Noise Model Estimation

As a first step, the model parameters of the channel-distorted noise signal $\mathbf{n_h}$ is estimated as a single Gaussian pdf $\{\boldsymbol{\mu}_{\mathbf{n_h}}, \boldsymbol{\Sigma}_{\mathbf{n_h}}\}$ in the cepstral domain. In this study, a cluster-based speech/non-speech segment detection method is employed, where input feature vectors are clustered to the two centroids by a simple binary splitting algorithm. Among the obtained two centroids, the one who has a lower value in the 0th cepstral component (i.e., lower energy) is decided as a centrod for the non-speech segments.

### 3.2. Step 2: Channel-Distored Speech Estimation

In this step, channel-distored speech signal $\mathbf{x_h}$ is estimated by the PCGMM-based feature compensation scheme [1]. Given an estimated channel distortion vector $\mathbf{h}$, the pdf of $\mathbf{x_h}$ can be represented as,

$$\{\omega_k, \boldsymbol{\mu}_{\mathbf{x_h},k}, \boldsymbol{\Sigma}_{\mathbf{x_h},k}\} = \{\omega_k, \boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{h}, \boldsymbol{\Sigma}_{\mathbf{x},k}\}, \qquad (2)$$

where $\{\omega_k, \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}\}$ is a set of the GMM parameters for clean speech $\mathbf{x}$ obtained by training over clean speech data. Here, the channel component $\mathbf{h}$ is assumed be additive to clean speech in the cepstral domain, so the mean parameter of the clean speech model is transformed by a bias term.

From the additivity of $\mathbf{x_h}$ and $\mathbf{n_h}$ in the waveform as given in Eq. (1), the model parameters of the noise-corrupted speech $\mathbf{y}$ are obtained by the model combination technique as follows:

$$\{\omega_k, \boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}\} = \mathcal{F}[\{\omega_k, \boldsymbol{\mu}_{\mathbf{x_h},k}, \boldsymbol{\Sigma}_{\mathbf{x_h},k}\}, \{\boldsymbol{\mu}_{\mathbf{n_h}}, \boldsymbol{\Sigma}_{\mathbf{n_h}}\}], \qquad (3)$$

where $\mathcal{F}[\cdot]$ denotes a function representing the model combination employed in the PCGMM method, and the same weight parameter $\omega_k$ is just used as in the clean speech model. In

this study, the log-normal approximation method is used for the model combination technique [1].

A constant bias transformation of the mean parameters of the channel-distorted speech model is assumed in the cepstral domain, which is the assumption generally taken by other data-driven methods [9] as follows,

$$\boldsymbol{\mu}_{\mathbf{y},k} = \boldsymbol{\mu}_{\mathbf{x_h},k} + \mathbf{r}_{\mathbf{h},k}. \quad (4)$$

The bias term $\mathbf{r}_{\mathbf{h},k}$ is used for reconstruction of the channel-distorted clean speech. The Minimum Mean Squared Error (MMSE) estimation equation for the channel-distorted speech estimation is approximated as follows [1][9],

$$\widetilde{\mathbf{x_h}}(t) = \int_{\mathcal{X}_h} \mathbf{x_h} p(\mathbf{x_h}|\mathbf{y}(t)) d\mathbf{x_h} \cong \mathbf{y}(t) - \sum_{k=1}^{K} \mathbf{r}_{\mathbf{h},k}\, p(k|\mathbf{y}(t)). \quad (5)$$

### 3.3. Step 3: Channel Distortion Estimation

The channel distortion vector $\mathbf{h}$ is estimated by employing the Expectation Maximization (EM) algorithm over the obtained $\mathbf{x_h}(t)$ as a similar manner suggested in [9]. The auxiliary function for the EM algorithm can be written as follows:

$$\begin{aligned} Q(\mathbf{h}, \tilde{\mathbf{h}}) &= E\{L(\mathbf{x_h}, \mathbf{s}|\tilde{\mathbf{h}})|\mathbf{x_h}, \mathbf{h}\} \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K} \frac{p(\mathbf{x_h}(t), k|\mathbf{h})}{p(\mathbf{x_h}(t)|\mathbf{h})} \log(p(\mathbf{x_h}(t), k|\tilde{\mathbf{h}})) \end{aligned} \quad (6)$$

Here, $(\mathbf{x_h}, \mathbf{s})$ constitutes "complete" data containing source information, that is for which Gaussian component generates $\mathbf{x_h}$. To find $\tilde{\mathbf{h}}$ that maximizes the auxiliary function, $\tilde{\mathbf{h}}$ satisfying $\nabla_{\tilde{\mathbf{h}}} Q(\mathbf{h}, \tilde{\mathbf{h}}) = 0$ is developed, which leads to the following solution:

$$\tilde{\mathbf{h}} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k|\mathbf{x_h}(t), \mathbf{h}) \boldsymbol{\Sigma}_{\mathbf{x},k}^{-1} (\mathbf{x_h}(t) - \boldsymbol{\mu}_{\mathbf{x},k})}{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k|\mathbf{x_h}(t), \mathbf{h}) \boldsymbol{\Sigma}_{\mathbf{x},k}^{-1}}, \quad (7)$$

where $\boldsymbol{\mu}_{\mathbf{x},k}$ and $\boldsymbol{\Sigma}_{\mathbf{x},k}$ are the model parameters of clean speech $\mathbf{x}$. The channel distortion vector is iteratively estimated through Steps 2 and 3, with an initial channel vector $\mathbf{h}_0$. The iteration stops when the likelihood score $\sum_{t=1}^{T} p(\mathbf{x_h}(t)|\tilde{\mathbf{h}})$ does not increase. In this study, zero vector is used for $\mathbf{h}_0$, and the maximum number of iterations is set to 10. The final estimate of the channel distortion vector is used for the clean speech reconstruction in the next step.

### 3.4. Step 4: Clean Speech Reconstruction

In this final step, in a similar manner as Step 2, the clean speech is reconstructed by the PCGMM method. The model parameters for noise-corrupted speech $\mathbf{y}$ are given by Eq. (3). Here another approximation of bias transformation in the mean parameters of the clean speech model is formulated in the cepstral domain as follows,

$$\boldsymbol{\mu}_{\mathbf{y},k} = \boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{r}_k. \quad (8)$$

The clean speech is reconstructed by the MMSE estimator as follows,

$$\widetilde{\mathbf{x}}(t) = \int_{\mathcal{X}} \mathbf{x} p(\mathbf{x}|\mathbf{y}(t)) d\mathbf{x} \cong \mathbf{y}(t) - \sum_{k=1}^{K} \mathbf{r}_k\, p(k|\mathbf{y}(t)). \quad (9)$$

Fig. 3 illustrates the block diagram of the proposed PCGMM-based feature compensation algorithm integrated with the iterative channel estimation method.

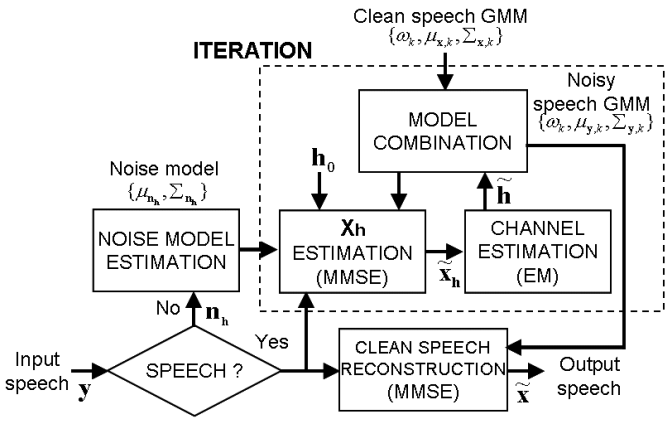


Figure 3: *Block diagram of the PCGMM-based feature compensation method integrated with iterative channel estimation.*

## 4. Experimental Results

### 4.1. Corpus Evaluation

To observe the degree of the noise/channel corruption of the speech corpus generated in this study, we evaluated several objective speech quality measures including SNR, PESQ [13], and speech recognition accuracy. The SNR was obtained using the NIST Speech Quality Assurance tool (i.e., the STNR estimator) [10]. Here the TIMIT+NOISE corpus represents a condition of additive background noise such as car and speech babble noise. The NTIMIT+NOISE and CTIMIT+NOISE corpora were obtained by the corpus generation procedure described in Sec. 2. They were generated by applying the channel distortion effects estimated from the NTIMIT and CTIMIT speech samples to the TIMIT+NOISE corpus respectively, formulating a presence of both additive and convolutional noise components, which is our interest in this study. From Tables 1 and 2, we can see that the STNR consistently decreases from clean condition to NTIMIT+NOISE and CTIMIT+NOISE. The PESQ also shows a similar trend to the STNR across all corpora.

Word Error Rate (WER) was examined as the speech recognition performance over each corpus. We employed SPHINX3 [11] as the Hidden Markov Model (HMM) based speech recognizer to obtain recognition performance. Each HMM represents a tri-phone which consists of 3 states with an 8-component GMM per state, which is tied with 1138 states. The task has 6233 words as the vocabulary, and the trigram language model is adapted on the TIMIT database using a Broadcast News language model as an initial model. A conventional MFCC feature front-end is employed in the experiment, which was suggested by the European Telecommunication Standards Institute (ETSI) [12]. An analysis window of 25 msec in duration is used with a 10 msec skip rate for 8-kHz speech data. The computed 23 Mel-filterbank outputs are transformed to 13 cepstrum coefficients including c0 (i.e., c0-c12).

Comparing the NTIMIT+NOISE to the TIMIT+NOISE, although the difference in the STNR is not large, the NTIMIT+NOISE shows significantly lower and higher values in PESQ and WER respectively. The CTIMIT+NOISE presents the lowest STNR and PESQ, and the highest WERs, which represent the most challenging condition for speech recognition. These results prove that the channel distortions applied to the TIMIT+NOISE bring significant corruption in signals, providing severely adverse environments for speech recognition.

Table 1: STNR (dB), PESQ (-0.5 to 4.5 MOS scale), and WER (%) over clean TIMIT, NTIMIT, CTIMIT, and TIMIT+NOISE corpora.

| | TIMIT Clean | NTIMIT | CTIMIT | TIMIT+NOISE | |
|---|---|---|---|---|---|
| | | | | Car | Babble |
| STNR | 54.40 | 35.79 | 24.14 | 16.90 | 17.58 |
| PESQ | 4.50 | 2.19 | 1.74 | 2.09 | 1.99 |
| WER | 8.05 | 34.33 | 76.19 | 62.36 | 51.34 |

Table 2: STNR (dB), PESQ (-0.5 to 4.5 MOS scale), and WER (%) over NTIMIT+NOISE and CTIMIT+NOISE corpora.

| | NTIMIT+NOISE | | CTIMIT+NOISE | |
|---|---|---|---|---|
| | Car | Babble | Car | Babble |
| STNR | 14.95 | 16.42 | 11.65 | 9.96 |
| PESQ | 1.69 | 1.64 | 1.55 | 1.52 |
| WER | 93.61 | 88.17 | 97.76 | 96.16 |

### 4.2. Performance Evaluation of the Proposed Feature Compensation Method

Performance of the proposed system (PCGMM+CH) was evaluated with comparison to several existing pre-processing algorithms in terms of speech recognition performance. Spectral Subtraction (SS) [14] combined with Cepstral Mean Normalization (CMN) was selected as one of the conventional algorithms. They represent some of the most commonly used techniques for additive noise suppression and removal of channel distortion respectively. We also evaluated the Vector Taylor Series (VTS) algorithm for performance comparison [9]. The Advanced Front-End (AFE) algorithm developed by ETSI was also evaluated as one of the state-of-the-art methods, which contains an iterative Wiener filter and blind equalization [15].

Table 3 shows speech recognition performance over the NTIMIT+NOISE and CTIMIT+NOISE corpora, using the proposed feature compensation method and existing pre-processing algorithms. The evaluation results over the NTIMIT+NOISE corpus indicate that the proposed PCGMM+CH shows slightly better performance in the average WER compared to the ETSI AFE. By combining CMN, the proposed method significantly outperforms the AFE[2]. The results for the CTIMIT+NOISE show that the WERs are around 60 % even for the ETSI AFE algorithm which is well known to be highly effective in noisy environment. This confirms that the CTIMIT+NOISE corpus has extremely challenging acoustic conditions for speech recognition. The proposed PCGMM+CH and its combination with CMN both significantly outperform the ETSI AFE with the relative improvements +8.11 % and +11.81 % respectively. The experimental results here demonstrate that the proposed PCGMM-based feature compensation method with channel estimation is highly effective in improving speech recognition performance in the severe adverse environments with the presence of both additive and convolutional noise components.

## 5. Conclusions

In this study, an effective feature compensation scheme was proposed to address severely adverse environment for speech recognition where background noise and channel distortion are simultaneously involved. The proposed scheme integrated an iterative channel estimation method into the framework of our PCGMM-based feature compensation algorithm. A new speech corpus was generated which reflects both additive and convolutional noise corruption. The channel distortion effects were obtained from the NTIMIT and CTIMIT corpora. Evaluation of objective measures including STNR, PESQ, and speech recognition showed that generated speech corpus includes highly challenging acoustic condition for speech recognition. Performance evaluation of the proposed system over the obtained speech corpus demonstrated that the proposed feature compensation scheme is significantly effective at improving speech recognition performance with presence of both background noise and channel distortion, comparing to the conventional methods including the ETSI AFE.

[2]ETSI AFE has shown the best performance when solely used without CMN over all corpora in this study.

Table 3: Recognition performance in WER (%) of the proposed system (PCGMM+CH) for NTIMIT+NOISE and CTIMIT+NOISE corpora with relative improvement to AFE.

| **NTIMIT + NOISE** | Car | Babble | Avg. | (Relative) |
|---|---|---|---|---|
| SS+CMN | 56.28 | 54.20 | 55.24 | |
| VTS | 56.28 | 49.21 | 52.70 | |
| ETSI AFE | 40.46 | 43.37 | 41.92 | |
| **PCGMM+CH** | **43.70** | **38.59** | **41.15** | **(+1.84)** |
| **PCGMM+CH+CMN** | **39.79** | **35.23** | **37.51** | **(+10.51)** |
| **CTIMIT + NOISE** | Car | Babble | Avg. | (Relative) |
| SS+CMN | 71.84 | 68.29 | 70.07 | |
| VTS | 72.65 | 66.74 | 69.70 | |
| ETSI AFE | 59.27 | 59.54 | 59.41 | |
| **PCGMM+CH** | **57.52** | **51.66** | **54.59** | **(+8.11)** |
| **PCGMM+CH+CMN** | **55.16** | **49.62** | **52.39** | **(+11.81)** |

## 6. References


[1] W. Kim and J.H.L. Hansen, "Feature Compensation in the Cepstral Domain Employing Model Combination," *Speech Comm.*, 51(2), pp.83-96, 2009.

[2] J.L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Proc.*, vol.2, no.2, pp.291-298, 1994.

[3] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," *Computer Speech and Language*, 9, pp.171-185, 1995.

[4] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," *IEEE Trans. on Speech and Audio Proc.*, vol.4, no.5, pp.352-359, 1996.

[5] H.G. Hirsch, and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," *ISCA ITRW ASR2000*, 2000.

[6] http://www.ldc.upenn.edu

[7] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database," *ICASSP-90*, pp. 109-112, 1990.

[8] K.L. Brown and E.B. George, "CTIMIT: A Speech Corpus for the Cellular Environment with Applications to Automatic Speech Recognition," *ICASSP-95*, pp. 105-108, 1995.

[9] P.J. Moreno, *Speech recognition in noisy environments*, Ph.D. Thesis. Carnegie Mellon University, 1996.

[10] http://www.nist.gov/speech.

[11] http://cmusphinx.sourceforge.net

[12] *ETSI standard document*, ETSI ES 201 108 v1.1.2 (2000-04), 2000.

[13] Y. Hu and P. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. on Speech and Audio Processing*, vol.16, no.1, pp.229-238, 2008.

[14] R. Martin, "Spectral Subtraction based on Minimum Statistics," . *EUSIPCO-94*, pp. 1182-1185, 1994.

[15] *ETSI standard document* ETSI ES 202 050 v1.1.1 (2002-10), 2002.