Chung and Hansen *EURASIP Journal on Audio, Speech, and Music Processing* 2013, **2013**:12
http://asmp.eurasipjournals.com/content/2013/1/12

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Compensation of SNR and noise type mismatch using an environmental sniffing based speech recognition solution

Yongjoo Chung[1*] and John HL Hansen[2]

Abstract

Multiple-model based speech recognition (MMSR) has been shown to be quite successful in noisy speech recognition. Since it employs multiple hidden Markov model (HMM) sets that correspond to various noise types and signal-to-noise ratio (SNR) values, the selected acoustic model can be closely matched with the test noisy speech, which leads to improved performance when compared with other state-of-the-art speech recognition systems that employ a single HMM set. However, as the number of HMM sets is usually limited due to practical considerations as well as effective model selection, acoustic mismatch can still be a problem in MMSR. In this study, we proposed methods to improve recognition performance by mitigating the mismatch in SNR and noise type for an MMSR solution. For the SNR mismatch, an optimal SNR mapping between the test noisy speech and the HMM was determined by experimental investigation. Improved performance was demonstrated by employing the SNR mapping instead of using the estimated SNR of the test noisy speech directly. We also proposed a novel method to reduce the effect of noise type mismatch by compensating the test noisy speech in the log-spectrum domain. We first derive the relation between the log-spectrum vectors in the test and training noisy speech. Since the relation is a non-linear function of the speech and noise parameters, the statistical information regarding the testing log-spectrum vectors was obtained by approximation using vector Taylor series (VTS) algorithm. Finally, the minimum mean square error estimation of the training log-spectrum vectors was used to reduce the mismatch between the training and test noisy speech. By employing the proposed methods in the MMSR framework, relative word error rate reduction of 18.7% and 21.3% was achieved on the Aurora 2 task when compared to a conventional MMSR and multi-condition training (MTR) method, respectively.

Keywords: Speech recognition; Multiple-model frame; Noise robustness; Environmental sniffing

1 Introduction

It is well known that significant performance degradation occurs when speech recognition is used in noisy environments. Various research efforts have previously been directed at noise-robust speech recognition such as noise-robust feature extraction, speech enhancement, and feature and model parameter compensation [1-5]. These approaches are used independently or in combination with each other to improve the performance of speech recognition under noisy environments.

Training hidden Markov model (HMM) directly using noisy speech has been considered an alternative approach to conventional methods [6-8]. However, such solutions work most effectively when the statistical structure of the noise does not vary greatly across train and test environments. Originally, developed to address speaking style/stress variations [6], multi-style training and later multi-condition training (MTR) have been considered on the Aurora 2 task [9]. In the MTR method, noisy speech signals under various noise conditions are used collectively for training the HMM. While remarkable performance improvements have been obtained with an MTR method, it has some drawbacks since it combines a number of noise conditions during training, which reduces the phonetic sharpness of the acoustic

* Correspondence: yjjung@kmu.ac.kr
[1]Department of Electronics, Keimyung University, 1000, Shindang-Dong, Dalseo-Gu, Daegu 704-701, South Korea
Full list of author information is available at the end of the article

models in their probability density functions versus matched training where the training material is assumed to have the same noise condition as the noisy test speech.

To overcome this weakness of the MTR method, a multiple-model based speech recognition (MMSR) framework was recently proposed, and successful results using this approach were demonstrated [8]. In this method, multiple HMM sets corresponding to various noise types and SNR values are constructed during training, and a single HMM (reference HMM) set which is closest to the noisy test speech is selected for recognition. MMSR has been shown to achieve better performance compared to MTR for the Aurora 2 task [8].

Before actual speech recognition takes place in the MMSR framework, it is first necessary to classify the noise type and estimate the SNR of the test speech in order to select the reference HMM that most closely matches the noise condition in the test speech. As errors in this process will cause misrecognition, performance of the MMSR can be improved significantly by minimizing or compensating for such errors. In the previous studies on MMSR, once the noise type is determined, the reference HMM that is closest to the estimated SNR of the test speech is selected [8]. However, according to our preliminary study [10], we expect that performance can be improved by selecting a reference HMM that has a slightly different (higher or lower) SNR value than the estimated SNR. This conjecture is based on our assumption that in a specific noise type/level, specific phoneme classes can be influenced by noise more than others (e.g., for wideband noise, consonants such as fricatives and stops will be more severely degraded, while vowels and diphthongs will have less distortion when considering the impact of noise on automatic speech recognition (ASR) performance). Also, speech signal energy is generally bimodal, with vowels, diphthongs, liquids, glides having high energy, and fricatives, stops, and affricates having low energy. The selection of an HMM with either higher or lower SNR may be influenced by the specific SNRs of consonants. Another possible reason for this SNR mismatch phenomenon was also explained in [11]. They say that training data with low SNR values reduces the speech discrimination of the trained HMM set, and it may be advantageous to employ an HMM set trained on data with higher SNR value.

According to a previous study [8], noise type classification accuracy using a Gaussian mixture model (GMM) for four known and distinct types of noise is nearly 100%. This suggests that noise type classification, assuming diverse and well-separated noise classes, does not adversely affect the performance of MMSR. However, this is no longer true when an unknown type of noise signal is present in the test speech. Since noise type mismatch can significantly impact performance, a strategy to address any noise type difference between training and test noisy speech should be employed. Since most conventional methods have focused on compensating the difference between clean and noise-corrupted speech, they cannot be directly applied to MMSR. Jacobian adaptation has been used to adapt the parameters of HMMs in changing noise conditions with some success [12,13]. However, since this is based on a simple linear approximation of the nonlinear cepstral distortion, it does not accurately reflect the changing noise conditions present in the HMM parameters.

Vector Taylor series (VTS) based approaches have been widely used for noise robust speech recognition [2,14] due to the outstanding performance of these methods. The basic strategy takes advantage of the relationship between clean and noise-corrupted speech signals in an analytical way where the relationship can be approximated quite accurately by the vector Taylor series. The resulting probability density function of the noisy speech signals can be easily estimated without using much adaptation data. Here, we apply a VTS-based approach to compensate for the noise type difference in MMSR. We first derive a novel formula describing the relationship between the test and training noisy speech in the log-spectrum domain and then VTS is used to approximate this nonlinear relation. During testing, we compensate the test log-spectrum vector to move it more closely towards a match with the reference HMM using minimum mean square error (MMSE) estimation of the training log-spectrum vector.

In this study, we propose to mitigate the mismatch between the test noisy speech and the selected reference HMM in MMSR from two different points of view. The SNR mismatch is reduced by optimally mapping the estimated SNR value of the test speech before selecting the reference HMM, and the noise type mismatch is handled by compensating the test noisy speech in the log-spectrum domain using VTS.

This paper is organized as follows: a review on MMSR is presented in section 2 and an experimental investigation on the SNR mismatch in the MMSR is described in section 3. Compensation of the test noisy speech is described in section 4. The experimental procedure and results are presented and discussed in section 5. Finally, conclusions are presented in section 6.

2 Multiple-model based speech recognition framework

2.1 Environmental sniffing

Environmental aware speech processing was proposed in the study by [15]. This previous study established the concept of 'environmental sniffing' in order to characterize and effectively direct subsequent speech processing systems based on environmental noise types and

levels. The study also showed that one could achieve a significant reduction in computational requirements versus traditional multi-recognizer ROVER solutions with an increase in recognition performance. Our study here focuses on selecting the best HMM platform from such an environmental sniffing hierarchy.

2.2 Architecture of multiple-model based speech recognition framework

In MMSR, multiple reference HMMs corresponding to the noise environments, both in type and SNR range, are constructed during training, and the reference HMM that is most appropriate for the test noisy speech is selected for recognition. To select the reference HMM, the SNR of the noisy test speech must be first estimated and the noise type classified. The architecture of the MMSR is shown in Figure 1.

In Figure 2, an example where a reference HMM is selected based on the SNR value and noise type of the test speech in the MMSR framework is shown. For this study, a reference HMM for every 2-dB interval across four noise types (babble, car, subway, exhibition) during training was constructed and stored in the environmental sniffing HMM database. In the example shown in Figure 2, the noise type from the noisy test speech was classified as subway and the SNR was estimated at 5.5 dB. This information was then sent to the environmental sniffing HMM database, and the reference HMM corresponding to subway/6 dB, which was closest to the noisy test speech, was selected for recognition. It is generally believed that choosing the reference HMM with the most similar SNR value to the test speech will result in the best ASR performance for conventional MMSR. However, in this study, we experimentally determined the optimal SNR value of the reference HMM which results in an improved recognition accuracy, better than matched, for a given noisy test speech utterance.

2.2.1 SNR estimation and noise type classification

A simple energy-based voice activity detector (VAD) [16] for SNR estimation in the MMSR was used. The VAD works in a similar way as an endpoint detector. It uses energy thresholds from the noisy speech utterance to find the speech parts of the utterance. In the SNR estimation method, the power of the noise $\hat{\sigma}_n^2$ was estimated using samples in the non-speech period obtained by the VAD, and this value was subtracted from the signal power $\hat{\sigma}_x^2$ estimated from the parts of speech activity to find the speech only power. The expression for the SNR in the noisy speech is defined as follows:

$$SNR = 10\log\frac{\hat{\sigma}_x^2 - \hat{\sigma}_n^2}{\hat{\sigma}_n^2} \tag{1}$$

For environmental sniffing based noise type classification, the cepstral feature vector of the noise signal **C**n is modeled by a GMM. The GMM represents the weighted linear combination of the Gaussian probability density functions and is expressed as follows:

$$\begin{aligned} p(\mathbf{C_n}) &= \sum_{i=1}^{M} \omega_i p_i(\mathbf{C_n}) \\ &= \sum_{i=1}^{M} \omega_i \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{C_n}-\boldsymbol{\mu}_i)'(\mathbf{\Sigma}_i)^{-1}(\mathbf{C_n}-\boldsymbol{\mu}_i)\right\} \end{aligned} \tag{2}$$

In Equation 2, the weight factor ω_i satisfies $\sum_{i=1}^{M} \omega_i = 1$ and $\boldsymbol{\mu}_i$, $\mathbf{\Sigma}_i$ each represent the D-dimensional mean vector and covariance matrix of the Gaussian probability density function $p_i(\mathbf{C_n})$. Other studies have also employed a GMM to characterize the acoustic noise space using online GMM modeling [17,18]. Next, for noise signal classification, the GMM is trained for each noise type via expectation maximization (EM)-based maximum-likelihood

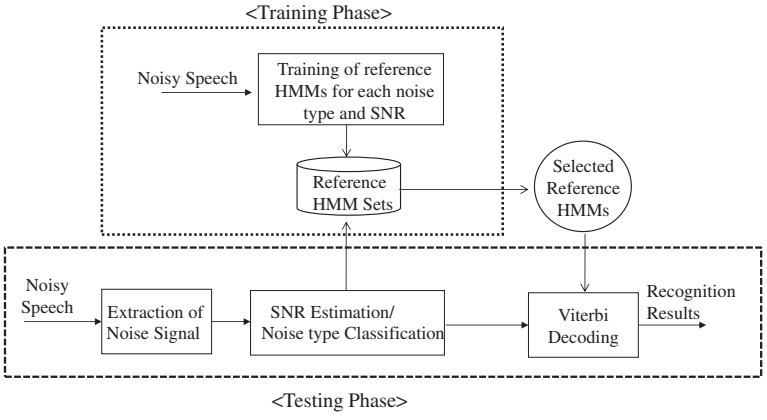

Figure 1 Architecture of MMSR framework which is divided into two parts: training phase and testing phase.

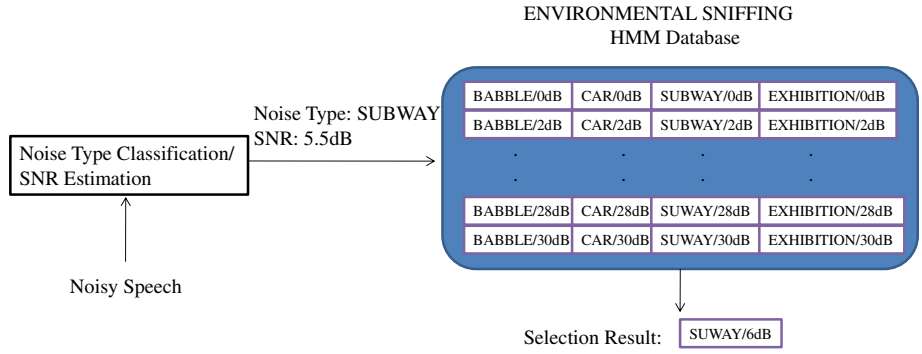

Figure 2 An example of the reference HMM selection process in MMSR.

estimation. Using the Aurora 2 database, one GMM is trained for each known noise type, and the GMM that provides the best accumulated likelihood for the first 10 frames of the noisy test speech is chosen as the noise type model for evaluation testing.

3 Analysis of SNR mismatch in environmental sniffing based MMSR

In this section, an experimental investigation was performed to illustrate the effect of SNR mismatch between noisy test and train speech in MMSR. The performance variation due to SNR mismatch was explored to determine the optimal SNR mapping between noisy test and train speech that gives the best recognition performance.

For the experiments in this section, a new train and test data was generated using the Aurora 2 database. Four known noise signal types (subway, babble, car, exhibition noises) were added to the clean training data of the Aurora 2 database to generate the noisy train data for each noise type across the SNR range from 0 to 30 dB in 2-dB intervals (16 SNR levels). The 4 × 16 sets of noisy training data were used to construct 64 sets of HMMs. The hidden Markov model toolkit [19] was employed for training and testing in this experiment. 1,001 clean speech sentences in Set A were corrupted to generate the noisy test data by adding the four known types of noise signal (subway, babble, car, exhibition noises) with a SNR range from 0 to 30 dB in 2-dB intervals. The test data was generated independently of the test data used in experiments in section 5 so that the analytical results in this section could be applied without loss of generality to the SNR mapping in the speech recognition experiments described in section 5. The range of noise types and the SNR values of the training and test speech were assumed to be known *a priori* in the MMSR.

To illustrate the impact of SNR mismatch between train and test for the environmental sniffing MMSR framework, Figure 3 shows the word error rate (WER) surface across the four noise types. In the analysis, the minimum WER did not necessarily occur when the SNR of the noisy test speech matched the training speech (i.e., if a match HMM were selected, the red minimum plots would all lie along the green diagonal of the input test speech SNR versus selected HMM SNR model). This means that the lowest WER cannot be guaranteed even if the SNR was matched. This may not be a serious issue for the recognition performance at high SNRs, since the WER surface is relatively flat in these regions. However, we can see that the WER surface has steep changes in slope at low SNR regions, which signifies the importance of finding an effective SNR mapping between training and test speech for optimal recognition performance.

Figure 4 shows the WER curve as the SNR of the selected reference HMM is changed when the SNR of test speech is fixed at 0, 2 and 4 dB (babble noise), respectively. For example, when the SNR of the test speech is 0 dB, the WER is 43.65% when the reference HMM trained at 0 dB was selected, while the WER decreased to 32.07% for the 6-dB reference HMM. This example is in contrast to the conventional idea that the best noisy speech recognition performance will be achieved when the SNRs of the test and train speech are matched. This performance difference is so large that an alternative selection process is needed for the reference HMM given the SNR of the test speech. In addition to that illustrated for babble noise, a similar result was also observed for the other three types of noise in Set A as well.

Based on the WER surface shown in Figure 3, it is clearly possible to determine the best reference HMM given the estimated SNR of the test speech. The results of this analysis are summarized in Table 1. As expected, some difference in SNR between the noisy test speech and the best reference HMM were observed. For the test speech with low SNRs, an advantage to selecting a reference HMM with a higher SNR value than the actual estimated value was demonstrated. In Table 1, the estimated SNR values of the test speech is adjusted to compensate

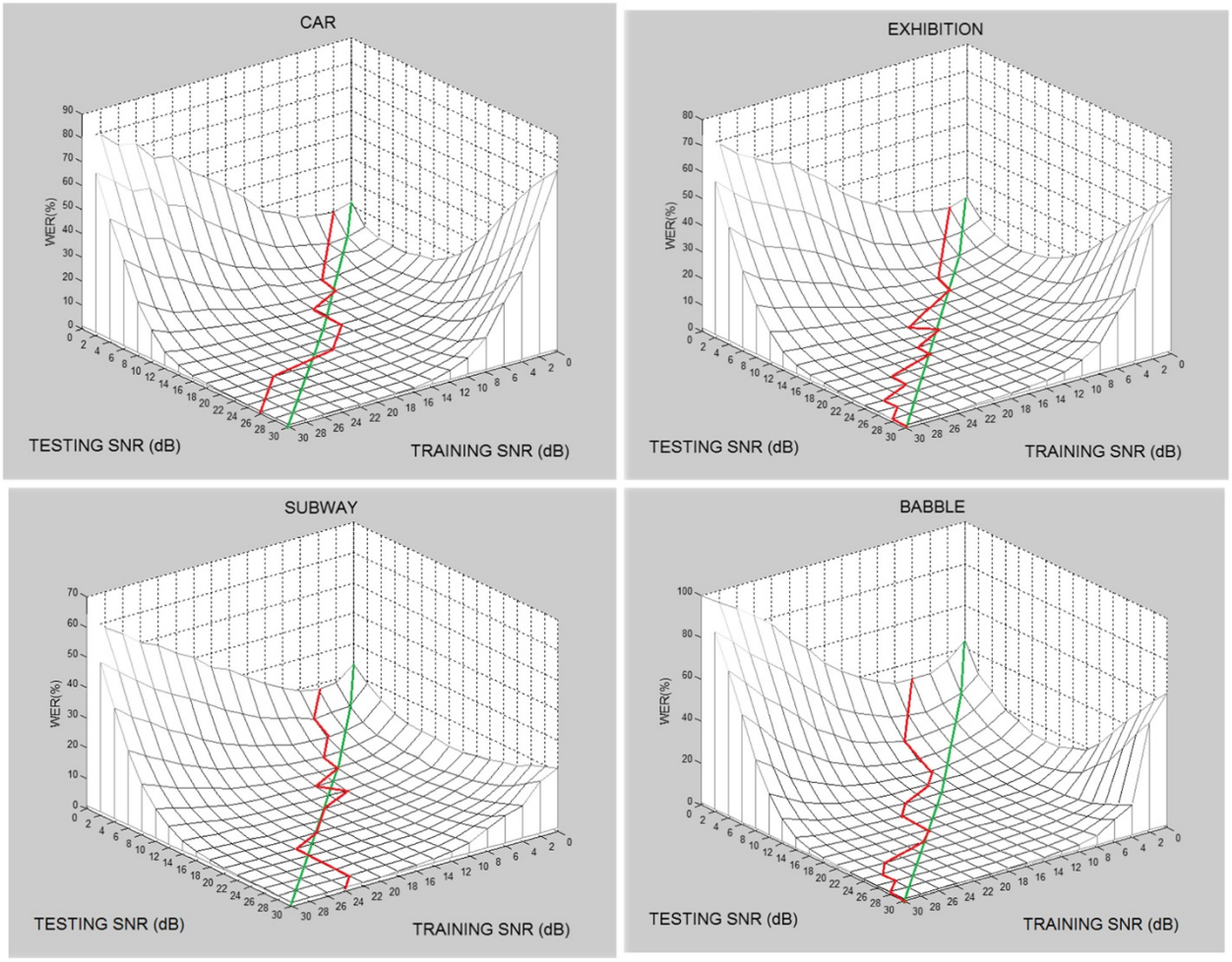

Figure 3 WER surface. Showing result of SNR mismatch between noisy test and train speech for four noise types in the Aurora 2 database (The red line connects the minimum WER points).

for the estimation errors, which makes the SNR mapping in Table 1 robust against the SNR estimation errors.

4 Feature compensation for environmental sniffing based MMSR

Although SNR mismatch in MMSR can be reduced through optimal mapping of the estimated SNR as described in section 3, there remains the problem of noise type mismatch, which occurs when an unknown (unseen during training) type of noise signal is present in the test speech. For this reason, the recognition accuracy of MMSR was worse than the MTR method for Set B (see section 5.1 for a detailed description on Set B) where unknown types of noise are encountered. To overcome this problem, we developed a novel feature compensation method based on VTS. First, the relation between the log-spectrum vectors in the noisy train and test speech was derived. Since this relation is a nonlinear function of the speech and noise parameters, the statistical information regarding the test log-spectrum vectors is obtained by approximation using the VTS algorithm. Finally, MMSE estimation for the training log-spectrum vectors is performed to reduce the mismatch between the noisy train and test speech.

4.1 Relationship between noisy speech signals

For feature compensation, a relationship between the log-spectrum vectors in the noisy training and test speech was derived. We employed the usual assumption of the relationship between the clean speech vector $\mathbf{x}$ and the noisy speech vector $\mathbf{y}$ in the log-spectrum domain as follows:

$$\mathbf{y} = \mathbf{x} + \log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{x})) \tag{3}$$

If $g(\mathbf{x}, \mathbf{n}) = \log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{x}))$, then we have $\mathbf{y} = \mathbf{x} + g(\mathbf{x}, \mathbf{n})$, where $\mathbf{n}$ is the additive noise in the corruption process and $\mathbf{i}$ is a unity vector. Based on Equation 3, the noisy

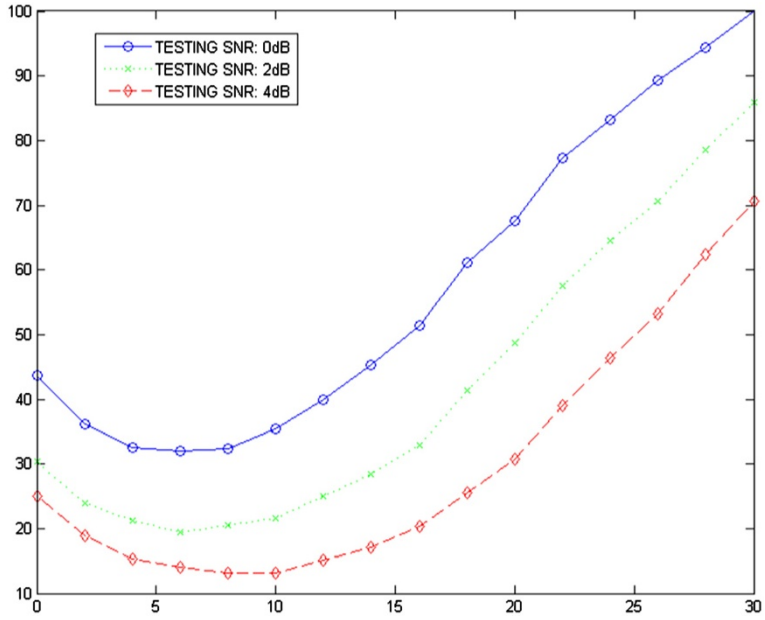

Figure 4 The variation of the WER (%) as the SNR of the reference HMM changes. SNR of the reference HMM changes when the SNR of the test speech is fixed at 0, 2, 4 dB, respectively.

log-spectrum vector $\mathbf{y}_{\mathrm{Tr}}$ in the training speech and $\mathbf{y}_{\mathrm{Te}}$ in the test speech can be described as follows:

$$\mathbf{y}_{\mathrm{Tr}} = \mathbf{x} + \log(\mathbf{i} + \exp(\mathbf{n}_{\mathrm{Tr}} - \mathbf{x})) = \mathbf{x} + g(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}}) \quad (4)$$

$$\mathbf{y}_{\mathrm{Te}} = \mathbf{x} + \log(\mathbf{i} + \exp(\mathbf{n}_{\mathrm{Te}} - \mathbf{x})) = \mathbf{x} + g(\mathbf{x}, \mathbf{n}_{\mathrm{Te}}), \quad (5)$$

where $\mathbf{n}_{\mathrm{Tr}}$ and $\mathbf{n}_{\mathrm{Te}}$ are the additive noises in the training and test speech, respectively.

Combining Equations 4 and 5, we can express $\mathbf{y}_{\mathrm{Te}}$ in terms of $\mathbf{y}_{\mathrm{Tr}}$ as follows:

$$\mathbf{y}_{\mathrm{Te}} = \mathbf{y}_{\mathrm{Tr}} + g(\mathbf{x}, \mathbf{n}_{\mathrm{Te}}) - g(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}}). \quad (6)$$

Assume that $\mathbf{n}_{\mathrm{Tr}}$ is determined beforehand during training and $\mathbf{n}_{\mathrm{Te}}$ is expressed as a variable $\mathbf{n}$, which should be estimated using the noisy test speech, then, $g(\mathbf{x}, \mathbf{n}_{\mathrm{Te}}) - g(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}})$ can be described as follows:

$$\begin{aligned} g(\mathbf{x}, \mathbf{n}_{\mathrm{Te}}) - g(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}}) &= \log(\mathbf{i} + \exp(\mathbf{n}_{\mathrm{Te}} - \mathbf{x})) \\ &\quad - \log(\mathbf{i} + \exp(\mathbf{n}_{\mathrm{Tr}} - \mathbf{x})) \end{aligned}$$

$$\begin{aligned} [g(\mathbf{x}, \mathbf{n}) - g(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}})]_i &= \log\left(\frac{[\mathbf{i} + \exp(\mathbf{n} - \mathbf{x})]_i}{[\mathbf{i} + \exp(\mathbf{n}_{\mathrm{Tr}} - \mathbf{x})]_i}\right) \\ &= \log\left(\frac{[\exp(\mathbf{x}) + \exp(\mathbf{n})]_i}{[\exp(\mathbf{x}) + \exp(\mathbf{n}_{\mathrm{Tr}})]_i}\right). \end{aligned} \quad (7)$$

Here, $[\cdot]_i$ represents the i^{th} element of a vector.

From Equation 4, the following equation can be derived:

$$\mathbf{y}_{Tr} = \log(\exp(\mathbf{x}) + \exp(\mathbf{n}_{Tr})) \quad (8)$$

Table 1 SNR of reference HMM showing lowest WER as the SNR of test speech was varied

SNR of test speech	SNR of reference HMM showing the lowest word error rate			
	Babble	Subway	Car	Exhibition
0	6	4	2	2
2	6	6	4	4
4	8	6	6	6
6	10	8	8	8
8	10	8	8	10
10	12	12	12	10
12	16	10	14	12
14	18	14	14	16
16	18	16	20	18
18	18	18	18	18
20	20	22	20	20
22	26	22	26	26
24	28	22	28	28
26	28	22	30	28
28	28	22	30	30
30	30	24	30	30

Taking the exponential of both sides in Equation 8, the equation can be rewritten as follows:

$$\begin{aligned} \exp(\mathbf{y}_{Tr}) &= \exp(\mathbf{x}) + \exp(\mathbf{n}_{Tr}) \\ \exp(\mathbf{x}) &= \exp(\mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr}) \end{aligned} \tag{9}$$

Substituting Equation 9 into Equation 7 produces

$$\begin{aligned} [g(\mathbf{x},\mathbf{n}) - g(\mathbf{x},\mathbf{n}_{Tr})]_i &= \log\left(\frac{[\exp(\mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr}) + \exp(\mathbf{n})]_i}{[\exp(\mathbf{y}_{Tr})]_i}\right) \\ &= [\log(\mathbf{i} + \exp(\mathbf{n}-\mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr}-\mathbf{y}_{Tr}))]_i \\ &\equiv [G(\mathbf{y}_{Tr},\mathbf{n},\mathbf{n}_{Tr})]_i \end{aligned} \tag{10}$$

If Equation 10 is substituted back into Equation 6, the relation between the log-spectrum vectors in the noisy training and test speech can be obtained as follows:

$$\begin{aligned} \mathbf{y}_{Te} &= \mathbf{y}_{Tr} + G(\mathbf{y}_{Tr},\mathbf{n},\mathbf{n}_{Tr}) \\ &= \mathbf{y}_{Tr} + \log(\mathbf{i} + \exp(\mathbf{n}-\mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr}-\mathbf{y}_{Tr})) \end{aligned} \tag{11}$$

Equation 11 can be used to find statistical information on $\mathbf{y}_{Te}$ given the statistics of the training log-spectrum vector $\mathbf{y}_{Tr}$.

4.2 Compensation of the feature vector

Next, a compensation of the feature vector was performed employing MMSE estimation of the log-spectrum vector. The compensation will estimate the noisy log-spectrum in the training speech given the log-spectrum from the test speech using a statistical relation. By using the estimated log-spectrum vector instead of the original log-spectrum vector from the test speech, the mismatch between the test speech and reference HMM in the environmental sniffing based MMSR (ESniff MMSR) can be reduced, which will improve the recognition performance without changing the parameters of the reference HMM.

4.2.1 Estimating the mean and covariance of log-spectrum vector

A statistical relationship between the log-spectrum vectors from the training and test noisy speech was first derived. Equation 11 was expanded by using the first-order VTS around an initial value $\mathbf{n}_0$ of $\mathbf{n}$, and the mean of the training log-spectrum vector $\boldsymbol{\mu}_{Tr} = E\{\mathbf{y}_{Tr}\}$ to obtain the following equation:

$$\begin{aligned} \mathbf{y}_{Te} = \mathrm{y}_{Tr} + G\left(\boldsymbol{\mu}_{\mathrm{y}_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right) + \nabla_{\mathbf{y}_{Tr}} G\left(\boldsymbol{\mu}_{y_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right)\left(\mathbf{y}_{Tr}-\boldsymbol{\mu}_{y_{Tr}}\right) + \\ \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right)(\mathbf{n}-\mathbf{n}_0) \end{aligned}, \tag{12}$$

where the gradient matrices are assumed to be diagonal and obtained as follows:

$$\begin{aligned} \left[\nabla_{\mathbf{y}_{Tr}} G\left(\boldsymbol{\mu}_{y_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right)\right]_{ii} &= \frac{[\exp(\mathbf{n}_{Tr}) - \exp(\mathbf{n}_0)]_i}{\left[\exp\left(\boldsymbol{\mu}_{y_{Tr}}\right) + \exp(\mathbf{n}_0) - \exp(\mathbf{n}_{Tr})\right]_i}, \\ \left[\nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right)\right]_{ii} &= \frac{[\exp(\mathbf{n}_0)]_i}{\left[\exp\left(\boldsymbol{\mu}_{y_{Tr}}\right) + \exp(\mathbf{n}_0) - \exp(\mathbf{n}_{Tr})\right]_i} \end{aligned} \tag{13}$$

Here, $[\cdot]_{ii}$ represents the i^{th} diagonal element of a matrix. Using Equation 12, the mean $\boldsymbol{\mu}_{y_{Te}}$ and covariance $\boldsymbol{\Sigma}_{y_{Te}}$ of $\mathbf{y}_{Te}$ can be expressed from the mean vector and covariance matrix of the noisy training speech $\mathbf{y}_{Tr}$ as follows:

$$\begin{aligned} \boldsymbol{\mu}_{y_{Te}} &= \boldsymbol{\mu}_{y_{Tr}} + G\left(\boldsymbol{\mu}_{y_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right) + \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right)(\mathbf{n}-\mathbf{n}_0) \\ \boldsymbol{\Sigma}_{y_{Te}} &= \left(\mathbf{I} + \nabla_{\mathbf{y}_{Tr}} G\left(\boldsymbol{\mu}_{y_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right)\right)\boldsymbol{\Sigma}_{y_{Tr}}\left(\mathbf{I} + \nabla_{\mathbf{y}_{Tr}} G\left(\boldsymbol{\mu}_{y_{Tr}},\mathbf{n}_0,\mathbf{n}_{Tr}\right)\right)^T, \end{aligned} \tag{14}$$

where $\mathbf{I}$ is an identity matrix. Next, the noise vector was characterized.

4.2.2. Maximum likelihood estimation of noise vector

The log-spectrum vector $\mathbf{y}_{Tr}$ of the noisy training speech was assumed to be distributed as a mixture of Gaussian distributions with mean vectors and covariance matrices obtained through a vector quantization process using the noisy training data. The mixture Gaussian distribution was separately estimated for each noisy type and SNR value using the same noisy training data to produce the reference HMM sets. Assuming also that the log-spectrum vector $\mathbf{y}_{Te}$ of the noisy test speech is a mixture of distributed Gaussians, the distribution of $\mathbf{y}_{Te}$ as a function of unknown noise vector $\mathbf{n}$ can be defined using Equation 14:

$$p(\mathbf{y}_{Te}|\ \mathbf{n}) = \sum_{m=1}^{M} p_m\ N\left(\boldsymbol{\mu}_{y_{Te},m},\ \boldsymbol{\Sigma}_{y_{Te},m}\right), \tag{15}$$

where $N\left(\boldsymbol{\mu}_{y_{Te},m},\ \boldsymbol{\Sigma}_{y_{Te},m}\right)$ is the m^{th} Gaussian component with a mean vector $\boldsymbol{\mu}_{y_{Te},m}$ and a covariance matrix $\boldsymbol{\Sigma}_{y_{Te},m}$. Also, p_m is the mixture weight of the m^{th} Gaussian component. Note that the mean vector $\boldsymbol{\mu}_{y_{Te},m}$ and covariance matrix $\boldsymbol{\Sigma}_{y_{Te},m}$ are, themselves, fully parameterized by the noise vector $\mathbf{n}$. In this study, the noise vector $\mathbf{n}$ was treated just as a parameter and not a random variable, and only the noisy speech vectors were treated as random variables.

Given a sequence of test log-spectrum vectors of length T, written as $\mathbf{Y}_{Te} = \{\mathbf{y}_{Te,1}, \mathbf{y}_{Te,2}, \cdots, \mathbf{y}_{Te,T}\}$, the resulting log-likelihood function is defined as follows:

$$L(\mathbf{Y}_{Te}|\mathbf{n}) = \sum_{t=1}^{T} \log\, p(\mathbf{y}_{Te,t}|\mathbf{n}). \tag{16}$$

Here, an iterative EM algorithm was employed to re-estimate the noise vector $\mathbf{n}$ by maximizing the log-likelihood function for the noisy test speech.

In the EM algorithm, the auxiliary function $Q(\boldsymbol{\varphi}, \bar{\boldsymbol{\varphi}})$ used is defined as follows:

$$\begin{aligned} Q(\boldsymbol{\varphi},\bar{\boldsymbol{\varphi}}) &= E\{L(\mathbf{Y}_{Te}|\bar{\boldsymbol{\varphi}})|\mathbf{Y}_{Te},\boldsymbol{\varphi}\} \\ &= \sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_{Te,t},\mathbf{n}) \log p(\mathbf{y}_{Te,t},m|\bar{\mathbf{n}}) \quad . \end{aligned}$$

The symbol $\boldsymbol{\varphi}$ actually represents the noise vector $\mathbf{n}$ which is assumed to be already known and $\bar{\boldsymbol{\varphi}}$ is the unknown noise vector $\bar{\mathbf{n}}$ which should be estimated. It is worth noting that $Q(\boldsymbol{\varphi}, \bar{\boldsymbol{\varphi}})$ can be expanded as follows:

$$\begin{aligned} Q(\boldsymbol{\varphi}, \bar{\boldsymbol{\varphi}}) = \sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_{Te,t},\mathbf{n}) \Bigg[&\log p_m + \frac{D}{2}\log 2\pi - \frac{D}{2}\log|\boldsymbol{\Sigma}_{y_{Te},m}| + \\ -\frac{1}{2}(\mathbf{y}_{Te,t} - (&\boldsymbol{\mu}_{y_{Tr},m} + G(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}) \\ &+ \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}\right)(\bar{\mathbf{n}} - \mathbf{n}_0))) T \boldsymbol{\Sigma}_{y_{Te},m}^{-1} \\ \cdot(\mathbf{y}_{Te,t} - (&\boldsymbol{\mu}_{y_{Tr},m} + G(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}) \\ &+ \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}\right)(\bar{\mathbf{n}} - \mathbf{n}_0))) \Bigg] \end{aligned} \tag{17}$$

Next, to re-estimate $\mathbf{n}$ in Equation 17, the derivative of the auxiliary function with respect to $\bar{\mathbf{n}}$ must be taken and set to equal 0.

$$\begin{aligned} \nabla_{\bar{\mathbf{n}}} Q(\boldsymbol{\varphi}, \bar{\boldsymbol{\varphi}}) = &\sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_{Te,t},\mathbf{n}) \\ &\Bigg[\nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{\mathrm{Tr}}\right)^{T} \boldsymbol{\Sigma}_{y_{Te},m}^{-1} \\ \cdot(\mathbf{y}_{Te,t} - (&\boldsymbol{\mu}_{y_{Tr},m} + G(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}, {}^{0}\mathbf{n}_{Tr}) \\ &+ \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}\right)(\bar{\mathbf{n}} - \mathbf{n}_0))) \Bigg] = 0 \end{aligned} \tag{18}$$

Table 2 Comparison of WERs (%) of ESniff MMSR with other approaches when using FE feature vectors

Approach	WER (%)			
	Set A	Set B	Set C	Average
CLEAN	38.66	44.25	33.86	39.94
VTS	28.23	29.31	24.95	28.00
PMC	20.70	18.82	21.98	20.20
MTR	12.23	13.75	16.42	13.68
ESniff MMSR	8.92	16.64	15.09	13.24

$$\begin{aligned} \bar{\boldsymbol{n}} = &\Bigg[\sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_{Te,t},\mathbf{n}) \;\; \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}\right)^{T} \\ &\boldsymbol{\Sigma}_{y_{Te},m} \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}\right)\Bigg]^{-1} \cdot \\ &\Bigg[\sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_{Te,t},\mathbf{n}) \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}\right)^{T} \boldsymbol{\Sigma}_{y_{Te},m}^{-1} \\ &\cdot(\mathbf{y}_{Te,t} - (\boldsymbol{\mu}_{y_{Tr},m} + G(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}, \mathbf{n}_{Tr}) \\ &\quad - \nabla_{\mathbf{n}} G\left(\boldsymbol{\mu}_{y_{Tr},m}, \mathbf{n}_0, \mathbf{n}_{Tr}\right)\mathbf{n}_0))]. \end{aligned} \tag{19}$$

The noise vector derived from Equation 19 was then substituted into Equation 14 to adapt $\boldsymbol{\mu}_{y_{Te},m}$ and $\boldsymbol{\Sigma}_{y_{Te},m}$ in Equation 15. The likelihood function from Equation 16 and the auxiliary function from Equation 17 were consequently updated. This process was iterated until a defined convergence criterion was met in the log-likelihood function from Equation 16 of the noisy test speech. After convergence, an MMSE estimation of the

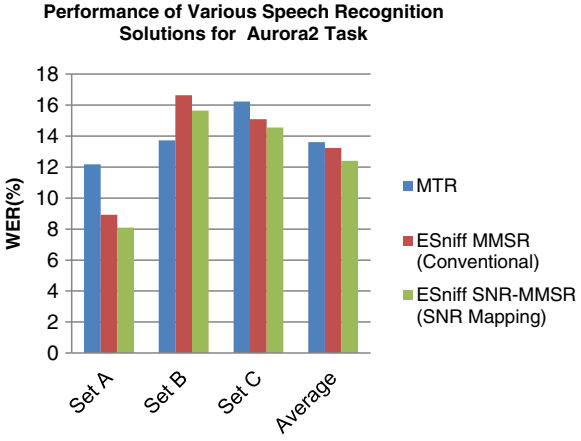

Figure 5 WERs (%) of the ESniff MMSR with SNR mapping. Compared with the conventional ESniff MMSR and MTR methods when using the FE feature vectors.

Table 3 WER (%) of MTR method for Aurora 2 task using the FE feature vectors

SNR (dB)	Set A					Set B					Set C		
	Noise type					Noise type					Noise type		
	Sub.	Bab.	Car	Exh.	Ave.	Res.	Str.	Air.	Sta.	Ave.	Sub.	Str.	Ave.
20	2.36	2.30	1.91	2.68	2.31	3.22	2.36	2.68	2.96	2.80	2.79	3.57	3.18
15	3.50	2.78	2.36	3.42	3.01	4.70	3.75	3.88	4.54	4.22	3.72	4.59	4.15
10	5.65	4.69	4.41	5.92	5.17	8.01	5.74	6.59	7.22	6.89	6.82	7.62	7.22
5	11.67	12.42	12.38	12.50	12.24	16.55	14.36	13.72	16.72	15.34	17.87	17.90	17.88
0	33.13	37.82	46.73	35.91	38.40	40.71	38.69	34.77	43.91	39.52	53.21	46.07	49.64
Ave.	11.26	12.00	13.56	12.09	12.23	14.64	12.98	12.33	15.07	13.75	16.88	15.95	16.42

Sub., Subway; Bab., babble; Exh., exhibition; Res., restaurant; Str., street; Air., airport; Sta., train station.

original noisy training speech $\mathbf{y}_{Tr}$ was performed using the statistical information from $\mathbf{y}_{Te}$. Using this MMSE process, the spectral mismatch between the noisy test speech and the selected reference HMM in the MMSR is expected to be reduced significantly.

4.2.3 MMSE estimation of the log-spectrum

The MMSE estimate of $\mathbf{y}_{\mathrm{Tr}}$ given $\mathbf{y}_{\mathrm{Te}}$ can be expressed as follows:

$$\hat{\mathbf{y}}_{Tr,MMSE} = E(\mathbf{y}_{Tr}|\mathbf{y}_{Te}) = \int \mathbf{y}_{Tr} p(\mathbf{y}_{Tr}|\mathbf{y}_{Te})\, d\mathbf{y}_{Tr} \quad . \tag{20}$$

From Equation 11,

$$\begin{aligned} \mathbf{y}_{Tr} &= \mathbf{y}_{Te} - \log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr} - \mathbf{y}_{Tr})) \\ &= \mathbf{y}_{Te} - G(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{n}_{Tr}). \end{aligned} \quad . \tag{21}$$

The following relationship was determined by substituting Equation 21 into Equation 20 and approximating $G(\mathbf{y}_{\mathrm{Tr}}, \mathbf{n}, \mathbf{n}_{\mathrm{Tr}})$ based on a Taylor series approximation of order zero around the mean value $\boldsymbol{\mu}_{y_{Tr},m}$

$$\begin{aligned} \hat{\mathbf{y}}_{Tr,MMSE} &= \mathbf{y}_{Te} - \int_{\mathbf{y}_{Tr}} G(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{n}_{Tr}) p(\mathbf{y}_{Tr}|\mathbf{y}_{Te}) \mathrm{d}\mathbf{y}_{\mathrm{Tr}} \\ &= \mathbf{y}_{Te} - \int_{\mathbf{y}_{Tr}} \sum_{m=1}^{M} G(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{n}_{Tr}) p(\mathbf{y}_{Tr}, m|\mathbf{y}_{Te}) d\mathbf{y}_{Tr} \\ &= \mathbf{y}_{Te} - \sum_{m=1}^{M} p(m|\mathbf{y}_{Te}) \int_{\mathbf{y}_{Tr}} G(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{n}_{Tr}) p(\mathbf{y}_{Tr}|m, \mathbf{y}_{Te}) d\mathbf{y}_{Tr} \\ &\approx \mathbf{y}_{Te} - \sum_{m=1}^{M} p\left(m\middle|\mathbf{y}_{Te}\right) G\left(\boldsymbol{\mu}_{Tr,\mathrm{m}}, \mathbf{n}, \mathbf{n}_{Tr}\right) \end{aligned} \tag{22}$$

The discrete cosine transform of the log-spectrum vector $\hat{\mathbf{y}}_{Tr,MMSE}$ in Equation 22 was performed to find a 13th order cepstrum vector. The c0 component in the cepstrum vector was replaced with the log-energy. The delta and acceleration (delta-delta) coefficients of the cepstrum vector were also calculated to produce a 39-dimensional enhanced feature vector, which was finally used for speech recognition evaluation.

5 Experiments and discussions

5.1 Baseline system and speech corpora employed

In this study, we employ the Aurora 2 database for experiments. There are two sets of training data, each corresponding to clean training (CLEAN) and multi-condition training (MTR). Each consists of 8,440 sentences. The MTR set consists of both clean and noisy

Table 4 WER (%) of conventional ESniff MMSR method for Aurora 2 task using FE feature vectors

SNR (dB)	Set A					Set B					Set C		
	Noise type					Noise type					Noise type		
	Sub.	Bab.	Car	Exh.	Ave.	Res.	Str.	Air.	Sta.	Ave.	Sub.	Str.	Ave.
20	1.69	1.81	1.58	2.04	1.78	1.93	2.54	2.42	2.56	2.36	2.09	2.75	2.42
15	2.43	2.27	1.91	2.72	2.33	4.70	3.87	4.03	4.66	4.31	3.59	4.75	4.17
10	4.08	4.44	3.52	4.94	4.24	9.52	8.13	7.46	8.98	8.52	5.93	9.79	7.86
5	7.58	12.76	7.99	9.60	9.48	24.69	20.62	17.89	18.98	20.54	13.48	24.26	18.97
0	22.63	41.48	20.19	22.74	26.76	54.90	49.82	42.35	42.83	47.47	30.06	54.02	42.04
Ave.	7.68	12.55	7.04	8.41	8.92	19.15	17.00	14.83	15.60	16.64	11.03	19.15	15.09

Table 5 WER (%) of ESniff SNR-MMSR method for Aurora 2 task using FE feature vectors

SNR (dB)	Set A					Set B					Set C		
	Noise type					Noise type					Noise type		
	Sub.	Bab.	Car	Exh.	Ave.	Res.	Str.	Air.	Sta.	Ave.	Sub.	Str.	Ave.
20	1.54	1.78	1.58	2.01	1.73	1.96	2.36	2.48	2.53	2.33	2.03	2.48	2.25
15	2.52	2.33	1.91	2.96	2.43	4.39	3.99	4.38	4.69	4.36	3.56	4.69	4.12
10	3.65	4.63	3.64	5.00	4.23	9.03	8.62	7.46	8.79	8.47	5.80	9.61	7.70
5	7.28	10.22	7.72	8.89	8.53	22.44	19.62	16.25	18.76	19.27	12.77	22.04	17.40
0	18.82	33.71	20.01	21.63	23.54	49.46	44.38	38.59	42.70	43.78	33.37	49.12	41.24
Ave.	6.76	10.53	6.97	8.10	8.09	17.46	15.79	13.83	15.49	15.64	11.51	17.59	14.55

speech signal that is artificially contaminated by various kinds (subway, car, exhibition, babble) of noise with SNR ranges from 0 to 20 dB in 5-dB intervals.

Recognition experiments were conducted on 3 test sets (Set A, Set B, Set C) that are corrupted by a range of noise types with a SNR range of 0, 5, 10, 15, 20 dB. For each noise type and SNR value, there are 1,001 sentences for recognition. Set A and Set B are corrupted by an additive noise distortion alone, and Set C is corrupted by a combination of convolution noise and additive noise.

Two widely known speech features were used for the experiments. The first, entitled FE, consists of 12^{th} order Mel-frequency cepstral coefficients, with the 0^{th} cepstral component set aside, which were appended with the log energy to form a 13^{th} order basic feature vector along with their delta and acceleration coefficients to construct a 39-dimensional feature vector for each frame [20]. The second feature set is a noise robust version of the FE, which is generally called advanced front-end (AFE) in the literature and is known to significantly reduce word error rates in noisy conditions [21]. Thirty-nine-dimensional feature vectors in the AFE that were consistent with the feature size used for the FE were constructed.

The HMM for each digit consists of 16 states with 3 Gaussian mixtures in each state. Silence is also modeled by a three-state HMM with six Gaussian mixtures in each state. Four known types of noise signal were added to the CLEAN training data to generate noisy speech for training the reference HMMs in the ESniff MMSR solution. To construct a sufficient number of reference HMMs, a noisy speech signal was generated for every 2-dB interval between 0 and 30 dB resulting in a collection of 16 reference HMM sets constructed for each noise type. The total number of reference HMM sets used in the experiment was $4 \times 16 = 64$, with a single HMM set selected for recognition depending on the noise type and SNR value of the noisy test speech.

5.2 Experimental results

5.2.1. Comparison with conventional methods

In Table 2, the WER of ESniff MMSR was compared with other approaches for noisy speech recognition using FE for feature vectors. From the table, it can be seen that ESniff MMSR significantly outperforms parallel model combination (PMC) [2] as well as the CLEAN training and VTS [3] method, but is only slightly better than the previous MTR method. Compared with ESniff MMSR, the MTR method shows strong noise robustness for Set B which consists of noisy speech corrupted with unknown types of noise signals (restaurant, street, airport, station). Even though the ESniff MMSR performs much better than the MTR method for Set A and Set C, the difference in the average WER between the ESniff MMSR and MTR is not significantly large (13.24% versus 13.68%) due to the results from Set B. The sharp probability density function of the acoustic model in ESniff MMSR seems to have adversely affected the speech recognition performance for Set B.

Figure 5 shows the WER of ESniff MMSR when the reference HMM was selected using the SNR mappings

Table 6 WERs (%) of ESniff MMSE

Framework	WER (%)			
	Set A	Set B	Set C	Average
MTR	*12.23*	*13.75*	*16.42*	*13.68*
ESniff MMSR (Conventional)	*8.92*	*16.64*	*15.09*	*13.24*
ESniff SNR-MMSR (SNR Mapping)	*8.09*	*15.64*	*14.55*	*12.40*
ESniff MMSE ($M = 128$)	9.41	15.58	13.11	12.61
ESniff MMSE ($M = 32$)	8.95	14.99	12.12	12.00
ESniff MMSE ($M = 16$)	8.81	14.48	13.24	11.96
ESniff MMSE ($M = 8$)	8.41	13.60	12.30	11.26
ESniff MMSE ($M = 4$)	*8.09*	*13.11*	*11.44*	*10.76*
ESniff MMSE ($M = 2$)	8.25	13.34	12.47	11.13

The number of Gaussian distributions varies with MMSE performed along with a comparison to MTR, conventional ESniff MMSR and ESniff SNR-MMSR when using the FE feature vectors.

Table 7 WER (%) of ESniff MMSE (M = 4) method for Aurora 2 task using FE feature vectors

SNR (dB)	Set A					Set B					Set C		
	Noise type					Noise type					Noise type		
	Sub.	Bab.	Car	Exh.	Ave.	Res.	Str.	Air.	Sta.	Ave.	Sub.	Str.	Ave.
20	1.54	1.65	1.61	1.98	1.69	1.61	2.38	1.67	2.11	1.94	2.16	2.64	2.40
15	2.65	2.51	1.94	2.88	2.49	3.73	3.43	2.73	3.97	3.46	3.12	4.68	3.90
10	3.59	4.40	3.77	5.51	4.32	7.92	6.51	5.81	7.20	6.86	5.65	7.61	6.63
5	7.03	11.11	7.46	8.57	8.54	19.89	13.98	13.14	16.32	15.83	11.17	15.62	13.39
0	18.95	34.57	19.46	20.58	23.39	47.02	33.33	31.50	37.86	37.43	26.51	35.26	30.88
Ave.	6.75	10.85	6.85	7.90	8.09	16.03	11.93	10.97	13.49	13.11	9.72	13.16	11.44

obtained from Table 1. For comparison, we also include the WER performance of conventional ESniff MMSR and the MTR methods. The results in this figure confirm the findings reported in Table 2 that the conventional ESniff MMSR performs better than the MTR method reducing the relative WER by 3.2%. The SNR mapping based ESniff MMSR method (ESniff SNR-MMSR) further improves the performance of the conventional ESniff MMSR. The ESniff SNR-MMSR method produces an average WER of 12.40% (95% confidence interval of the WER is ±0.095%), thereby reducing the average relative WER by 6.3% and 9.4% compared with conventional ESniff MMSR and MTR methods, respectively. As shown in Figure 5, the ESniff SNR-MMSR performs better than the conventional ESniff MMSR for all three test sets (Set A, Set B, Set C), which demonstrates that the experimentally motivated SNR mappings in Table 1 are quite effective irrespective of the noise type in the test speech. Although SNR mapping has been established using the known types of noise signal during training, it was also found to be effective for the unknown types of additive noise signal in Set B and the convolution noise in Set C. More detailed results on the MTR, ESniff MMSR and ESniff SNR-MMSR can be found in Tables 3, 4 and 5, for the individual noise types from the Aurora 2 task.

To effectively address the problem of noise type mismatch, the MMSE estimation of the log-spectrum vectors given the noisy test speech was performed as described in section 4. The experimental results of this analysis demonstrate that the recognition performance of the proposed environmental sniffing based MMSE (ESniff MMSE) method depends on the number of Gaussian distributions in Equation 15. Table 6 shows the WER of the ESniff MMSE method as the number of Gaussian distributions is varied from 2 to 128. The SNR mapping was also applied to the ESniff MMSE. Table 6 shows that the average WER consistently drops as the number of Gaussian distributions is decreased from 128 to 4. The worst performance was observed when M = 128 and the best was obtained when M = 4. This means that a small number of Gaussian mixtures is more than adequate to model the noisy log-spectrum vectors. A small number of Gaussian mixtures may be more appropriate for the noisy speech signal which is spectrally flattened due to the high amplitude noise signal at low SNRs thereby eliminating the adverse effect of a poor fit to the test data at low SNRs. More detailed results

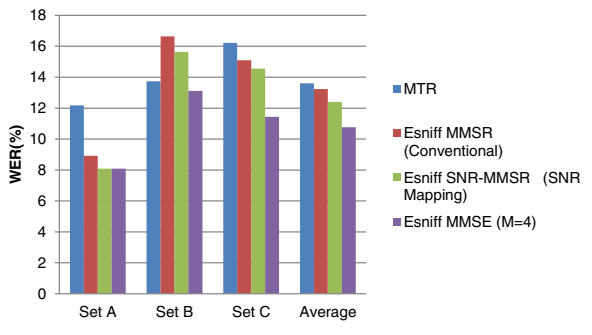

Figure 6 WERs (%) of ESniff SNR-MMSR and ESniff MMSE. In comparison with conventional ESniff MMSR and MTR methods when using FE feature vectors.

Table 8 WERs (%) of ESniff MMSE

Framework	WER (%)			
	Set A	Set B	Set C	Average
MTR	*7.70*	*8.23*	*9.26*	*8.22*
ESniff MMSR (Conventional)	7.59	10.66	8.57	9.01
ESniff SNR-MMSR (SNR Mapping)	6.78	9.56	8.17	8.17
ESniff MMSE (M = 128)	6.71	8.98	7.92	7.86
ESniff MMSE (M = 32)	6.69	9.01	8.00	7.88
ESniff MMSE (M = 16)	6.68	9.02	8.02	7.88
ESniff MMSE (M = 8)	6.61	8.97	8.01	7.86
ESniff MMSE (M = 4)	6.63	9.02	8.03	7.86
ESniff MMSE (M = 2)	8.01	10.58	9.56	9.34

The number of Gaussian distributions varies with MMSE performed along with a comparison to MTR, conventional ESniff MMSR and ESniff SNR-MMSR when using the AFE feature vectors.

Table 9 WER (%) of MTR method for Aurora 2 task using AFE feature vectors

SNR (dB)	Set A					Set B					Set C		
	Noise type					Noise type					Noise type		
	Sub.	Bab.	Car	Exh.	Ave.	Res.	Str.	Air.	Sta.	Ave.	Sub.	Str.	Ave.
20	1.26	1.45	1.16	1.48	1.34	1.57	1.75	1.13	1.05	1.37	1.54	1.72	1.63
15	2.33	2.27	1.49	2.04	2.03	2.24	2.33	2.03	1.88	2.12	2.18	2.39	2.28
10	4.73	3.96	2.56	4.04	3.82	4.51	3.84	3.67	3.42	3.86	3.87	4.38	4.12
5	8.41	9.67	6.23	8.76	8.27	11.42	9.61	7.46	9.04	9.38	9.46	10.46	9.96
0	22.44	29.90	19.21	22.71	23.06	29.72	23.40	21.32	23.14	24.39	27.72	28.84	28.28
Ave.	9.83	9.45	5.73	9.81	7.70	9.89	8.19	7.12	7.71	8.23	8.95	9.56	9.26

employing the ESniff MMSE solution is shown in Table 7. For the results in Table 7 and Figure 6, we selected M = 4 which gave the best performance on the Set A test.

Compared with the ESniff SNR-MMSR, ESniff MMSE always outperforms based on average WER, except when the MMSE Gaussian set is M = 128. The ESniff MMSE method generally shows significant performance improvements for both Set B and Set C, but the recognition accuracy is lower for Set A except when M = 4. This is expected since the ESniff MMSE was proposed to mitigate the effect of noise difference between noisy test and training speech. When the number of Gaussian distributions for ESniff MMSE is not appropriate, this approach adversely affects recognition performance for Set A which consists of noisy speech signals with *known types of noise* that do not require additional compensation for the noise type difference.

Figure 6 shows WER for the proposed methods (ESniff MMSE and ESniff SNR-MMSR) and compares this with MTR and conventional ESniff MMSR. The performance of MTR has been considered a benchmark in noisy speech recognition for the Aurora 2 task. As shown in the figure, the performance improvement for conventional ESniff MMSR is not significant compared with the MTR method. The ESniff SNR-MMSR method does improve the performance of the conventional ESniff MMSR by a significant margin, but it has limited performance benefit versus the MTR method for Set B due to the noise type mismatch between training and test speech. The decrease in performance due to the noise type mismatch could be significantly reduced by employing the ESniff MMSE method. By choosing the number of Gaussian distributions to be less than 8, better recognition performance is achieved versus MTR for Set B. When the number of Gaussian distributions is 4, the best average WER of 10.76% is obtained which corresponds to a reduction in relative WER of the MTR method by +21.3%. This performance improvement is significant compared with conventional ESniff MMSR where the relative WER is reduced by only +3.2% compared to the MTR method. Thus, in this study, speech recognition accuracy far better than conventional ESniff MMSR as well as the MTR method is achieved by employing the SNR mapping and MMSE estimation of the log-spectrum vector. This was achieved within an environmental sniffing framework, illustrating that effective SNR estimation with an improved mapping selection results in improved HMM speech recognition in noisy environments.

5.2.2 Performance evaluation using AFE feature vectors

In Table 8, the performance of the proposed methods (ESniff MMSE and ESniff SNR-MMSR) is compared with the MTR and conventional ESniff MMSR when using AFE feature vectors. Compared to conventional

Table 10 WER (%) of conventional ESniff MMSR method for Aurora 2 task using AFE feature vectors

SNR (dB)	Set A					Set B					Set C		
	Noise type					Noise type					Noise type		
	Sub.	Bab.	Car	Exh.	Ave.	Res.	Str.	Air.	Sta.	Ave.	Sub.	Str.	Ave.
20	1.29	1.45	1.07	1.36	1.29	1.50	2.12	1.10	1.23	1.49	1.47	2.03	1.75
15	2.06	2.09	1.22	1.79	1.79	2.49	2.81	1.79	2.62	2.43	2.18	2.60	2.39
10	4.11	4.47	2.36	3.67	3.65	5.04	5.71	4.41	4.57	4.93	3.99	5.11	4.55
5	7.61	11.61	5.10	7.99	8.08	15.11	13.88	10.20	12.03	12.80	8.63	11.67	10.15
0	20.85	37.24	14.17	20.27	23.13	38.56	33.59	28.00	26.50	31.66	20.91	27.09	24.00
Ave.	7.18	11.37	4.78	7.02	7.59	12.54	11.62	9.10	9.39	10.66	7.44	9.70	8.57

Table 11 WER (%) of ESniff SNR-MMSR method for Aurora 2 task using AFE feature vectors

SNR (dB)	Set A					Set B					Set C		
	Noise type					Noise type					Noise type		
	Sub.	Bab.	Car	Exh.	Ave.	Res.	Str.	Air.	Sta.	Ave.	Sub.	Str.	Ave.
20	1.29	1.18	1.04	1.20	1.18	1.54	2.06	1.22	1.05	1.47	1.38	1.93	1.65
15	1.90	2.15	1.16	1.82	1.76	2.61	2.99	2.06	2.53	2.55	2.15	2.69	2.42
10	3.87	3.96	2.36	3.76	3.49	4.54	5.96	4.26	4.57	4.83	3.81	5.32	4.56
5	7.43	9.55	5.07	8.11	7.54	13.14	12.88	9.22	10.49	11.43	7.74	10.88	9.31
0	18.02	20.44	13.51	18.82	19.95	32.55	29.35	23.41	24.84	27.54	19.59	26.24	22.91
Ave.	6.50	9.26	4.63	6.74	6.78	10.88	10.65	8.03	9.70	9.56	6.93	9.41	8.17

ESniff MMSR, ESniff SNR-MMSR shows a significant performance improvement as expected, and this improvement is consistently seen for all three test sets (Set A, Set B, and Set C). This demonstrates that the SNR mapping is still effective and not as sensitive to the specific feature vectors used. By employing ESniff MMSE, further improvements in performance based on the average WER were obtained. The performance of ESniff MMSE is robust against the change in the number of Gaussian distributions. The average WER of the ESniff MMSE varies only slightly as the number of Gaussian distributions is decreased from 128 to 4. This is in contrast to the results presented in Table 6 where a significant performance variation was observed with the number of Gaussian distributions. This may be due to the fact that the speech enhancement algorithm within the AFE has greatly reduced the noise signal in the test speech, and thus, a small number of Gaussian distributions are not as necessary.

When using AFE for feature vectors, the performance of conventional ESniff MMSR is found to be inferior to MTR, but both proposed methods (ESniff SNR-MMSR and ESniff MMSE) show improved results over MTR. However, the performance improvement is not as significant as when using FE feature vectors and MTR still performs better than the proposed methods for Set B. The use of speech enhancement algorithm within the AFE seems to reduce the relative improvement of the proposed methods over MTR. More detailed results on the MTR, ESniff MMSR, ESniff SNR-MMSR, and ESniff MMSE (M = 128) can be found in Tables 9, 10, 11, and 12.

6 Conclusions

This study demonstrated that an environmental sniffing based MMSR solution improves ASR performance over the conventional MTR method. However, the mismatch in noise type and SNR value between test and training speech make it difficult for the MMSR to perform significantly better than the MTR method. In this study, we developed methods to improve the performance of the conventional MMSR by reducing mismatch issues for noisy speech recognition within an environmental sniffing framework.

For the SNR value mismatch, we experimentally determined the SNR mappings between the noisy test and training speech for optimal recognition performance. We achieved an average WER of 12.40% on the Aurora 2 task using FE for feature vectors thereby reducing the average relative WER by 6.3% and 9.4% compared with conventional MMSR and MTR methods, respectively. This is remarkable considering the fact that the conventional MMSR method could reduce relative WER by just 3.2% compared to the MTR method. Although the SNR mapping was determined using training data with known types of noise signal, it was shown to possess a generalization property that improved recognition

Table 12 WER (%) of ESniff MMSE (M = 128) method for Aurora 2 task using AFE feature vectors

SNR (dB)	Set A					Set B					Set C		
	Noise type					Noise type					Noise type		
	Sub.	Bab.	Car	Exh.	Ave.	Res.	Str.	Air.	Sta.	Ave.	Sub.	Str.	Ave.
20	1.20	1.15	1.07	1.23	1.16	1.47	2.00	1.10	1.02	1.40	1.44	1.81	1.62
15	2.03	2.18	1.19	1.82	1.80	2.43	2.96	1.85	2.34	2.39	2.27	2.51	2.39
10	3.90	3.84	2.45	3.92	3.53	4.21	5.96	3.70	4.04	4.48	3.72	4.78	4.25
5	7.58	8.98	5.16	7.84	7.39	12.22	12.30	8.47	9.81	10.70	9.71	10.76	9.23
0	17.93	27.90	13.54	19.35	19.68	30.18	28.33	21.83	23.48	25.95	19.71	24.49	22.10
Ave.	6.53	8.81	4.68	6.83	6.71	10.10	10.31	7.39	8.14	8.98	6.97	8.87	9.92

performance on noisy test speech with combined unknown types of both additive and convolution noises.

The SNR mapping method improved performance over conventional MMSR by a significant margin but its performance is still inferior to MTR for Set B due to the noise type mismatch between training and test speech. It was possible to overcome this issue by MMSE of the training noisy log-spectrum given the test noisy speech. The performance of the MMSE method was found to be dependent on the number of Gaussian mixtures which model the noisy log-spectrum vectors. Compared with MTR and the conventional MMSR method, this solution showed improved performance for a wide range of Gaussian mixture counts. In particular, a small number of Gaussian mixtures was found to be more adequate in modeling the noisy log-spectrum vectors. As expected, the performance improvement was prominent for the test set with unknown types of additive noise signal. The MMSE method combined with the SNR mapping could reduce relative WER of the MTR method by 21.3% when using FE for feature vectors. This performance improvement is quite remarkable compared with the conventional MMSR method. When employing the AFE feature vectors, an improvement in performance was also observed using the proposed methods in noisy speech recognition, although the relative improvement over conventional methods was somewhat reduced due to the integrated speech enhancement algorithm inherent in the AFE.

In this study, we employed the SNR mapping and MMSE of the log-spectrum vectors in the environmental sniffing-based MMSR in an innovative way and achieved measurably improved speech recognition accuracy versus conventional MMSR as well as MTR methods. The results of this study show that an effective environmental sniffing framework coupled with improved SNR estimation and mapping, along with advanced noise modeling can improve overall speech recognition robustness.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011–0006994).

Author details
[1]Department of Electronics, Keimyung University, 1000, Shindang-Dong, Dalseo-Gu, Daegu 704-701, South Korea. [2]Center for Robust Speech System (CRSS), Eric Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA.

Received: 21 November 2012 Accepted: 6 June 2013
Published: 20 June 2013

References
1. S.F. Ball, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust., Speech, Signal Process **27**(2), 113–120 (1979)
2. M.J.F. Gales, *Model based techniques for noise-robust speech recognition, Dissertation* (University of Cambridge, Cambridge, 1996)
3. P.J. Moreno, *Speech Recognition in noisy environments, Dissertation* (Carnegie Mellon University, USA, 1996)
4. J.H.L. Hansen, Clements, A Mark, Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. on Signal Processing **39**(4), 795–805 (1991)
5. W. Kim, J.H.L. Hansen, Feature compensation in the cepstral domain employing model combination. Speech Commun. **51**(2), 83–96 (2009)
6. R.P. Lippmann, E.A. Martin, D.B. Paul, *Multi-style training for robust isolated-word speech recognition* (Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1987), Dallas, TX, 1987), pp. 705–708
7. Y. Gong, Speech recognition in noisy environments: a survey. Speech Commun. **16**, 261–291 (1995)
8. H. Xu, Z.H. Tan, P. Dalsgaard, B. Lindberg, *Robust speech recognition on noise and SNR classification – a multiple-model framework* (Proceedings of the 6th Annual Conference of the Speech Communication Association (INTERSPEECH 2005), Lisboa, Portugal, 2005), pp. 977–980
9. H.G. Hirsch, D. Pearce, *The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions* (Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000), Bejing, China, 2000), pp. 18–20
10. Y.J. Chung, *Optimal SNR model selection in multiple-based speech recognition system* (Proceedings of the First International Conference of Engineering and Technology Innovation (ICETI 2011), Kenting, Taiwan, 2011), pp. 154–159
11. H. Xu, X.H. Tan, P. Dalsgaard, B. Lindberg, Noise condition-dependent training based on noise classification and SNR estimation. IEEE Trans. Audio, Speech, Language Process **15**(8), 2431–2443 (2007)
12. S. Sagayama, Y. Yamaguchi, S. Takahashi, *Jacobian adaptation of noisy speech models* (Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 1997), Santa Barbara, California, 1997), pp. 396–403
13. R. Sarikaya, J.H.L. Hansen, *Improved Jacobian adaptation for fast acoustic model adaptation in noisy speech recognition* (Proceedings of the 1st Annual Conference of the Speech Communication Association (INTERSPEECH 2000), Beijing, China, 2000), pp. 702–705
14. D.Y. Kim, C.K. Un, N.S. Kim, Speech recognition in noisy environments using first-order vector Taylor series. Speech Communication **24**(1), 39–49 (1998)
15. M. Akbacak, J.H.L. Hansen, *Environmental sniffing: noise knowledge estimation for robust speech systems* (Proceedings of IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP 2003), Hongkong, 2003), pp. 113–116
16. L. Lamel, L. Rabiner, A. Rosenberg, J. Wilpon, An improved endpoint detector for isolated word recognition. IEEE Trans. Acoust., Speech. Signal Process **29**(4), 777–785 (1981)
17. M. Akbacak, J.H.L. Hansen, Environmental sniffing: noise knowledge estimation for robust speech systems. IEEE Trans. Audio, Speech and Language Process **15**(2), 465–477 (2007)
18. W. Kim, J.H.L. Hansen, A. Novel, Mask Estimation Method Employing Posterior-Based Representative Mean Estimate for Missing-Feature Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing **19**(5), 1434–1443 (2011)
19. S. Young, *HTK: Hidden Markov Model Toolkit V3.4.1* (Cambridge Univ. Eng. Dept. Speech Group, Cambridge, 1993)
20. ETSI draft standard doc, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithm. ETSI Standard ES 202 108*, 2000
21. ETSI draft standard doc, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithm. ETSI Standard ES 202 050*, 2002

doi:10.1186/1687-4722-2013-12
Cite this article as: Chung and Hansen: **Compensation of SNR and noise type mismatch using an environmental sniffing based speech recognition solution.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:12.