Noise Robust Speech Recognition Based on Noise-adapted HMMs Using Speech Feature Compensation

Yong-Joo Chung

Keimyung University, Daegu, S. Korea

yjjung@kmu.ac.kr

Abstract

In conventional VTS-based noisy speech recognition methods, the parameters of the clean speech HMM are adapted to test noisy speech, or the original clean speech is estimated from the test noisy speech. However, in noisy speech recognition, improved performance is generally expected by employing noisy acoustic models produced by methods such as Multi-condition TRaining (MTR) and Multi-Model based Speech Recognition (MMSR) framework compared with using clean HMMs. Motivated by this idea, a method has been developed that can make use of the noisy acoustic models in the VTS algorithm where additive noise was adapted for the speech feature compensation. In this paper, we modified the previous method to adapt channel noise as well as additive noise. The proposed method was applied to noise-adapted HMMs trained by the MTR and MMSR and could reduce the relative word error rate by 6.5% and 7.2%, respectively, in the noisy speech recognition experiments on the Aurora 2 database.

Index Terms: noise robust speech recognition, VTS, HMM

1. Introduction

Despite many technical advances, accurate speech recognition in noisy environments still remains a difficult problem. The techniques cannot fully overcome the performance degradation caused by channel and additive noise. Broadly categorized, there are two different approaches used to improve the performance in noisy speech recognition. In one of the approaches, test noisy speech or trained acoustic models are compensated to reduce the mismatch between them [1-4]. In particular, compensation based on Vector Taylor Series (VTS) has been known to perform quite well in noisy conditions [3,4].

In another approach, noisy speech was directly used to produce noise-adapted hidden Markov Models (HMMs) during training [5-7]. MTR [8] and MMSR [9,10] are representatives of this approach. The environment-dependent HMMs make it possible to cope with test noisy speech without any compensation algorithm.

Although the noise-adapted HMM performs rather well by itself, its performance would be improved further by applying compensation. In a previous study, a novel mathematical relation between test and training noisy speech was derived in the log-spectrum domain [11]. After approximating the relation using VTS, the performance of the noise-adapted HMM could be improved by compensating the feature vectors of the test noisy speech. However, in the previous study, the channel noise was not considered in the compensation, which probably had a negative effect on improving the performance on Set C of the Aurora 2 database. In this study, the previous algorithm was modified to compensate the test noisy speech considering both the channel and additive noise. The detailed mathematical formulation is derived, and MTR as well as MMSR are used for producing the noise-adapted HMM.

This paper is organized as follows. A review on MTR and MMSR is presented in Section 2, and compensation of the test noisy speech based on the noise-adapted HMM is described in Section 3. The experimental procedure and results are presented and discussed in Section 4. Finally, conclusions are given in Section 5.

2. A Review on Noise Adapted HMMs

In this study, both MTR and MMSR are used to produce the noisy speech HMM. Although MMSR is known to have some advantages over the MTR method [9,10], it is rather controversial regarding which method is better in performance for noisy speech recognition. In this regard, both methods will be used to find the more appropriate one in the proposed feature-compensation method.

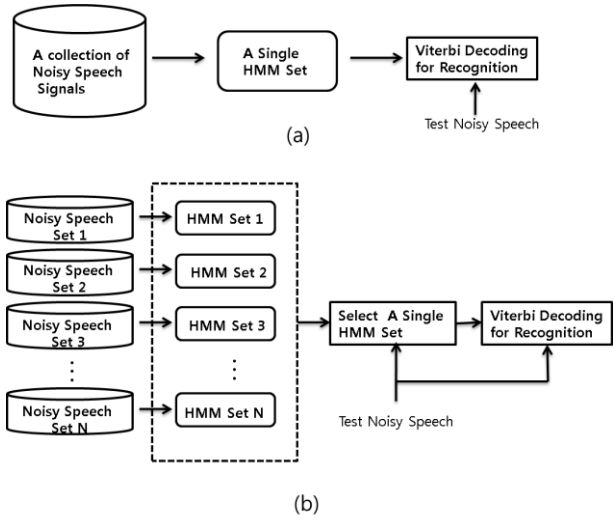

Figure 1. A Schematic diagram of training noisy speech HMMs (a) MTR (b) MMSR

In Figure 1, a schematic diagram of MTR and MMSR for training noisy speech HMMs is shown. In MTR, a collection of noisy speech signals with various noise types (Subway, Babble, Car, Exhibition) and SNR values (0, 5, 10, 15, 20 and ∞ dB) is used to construct a single set of noisy speech HMM. In MMSR, multiple HMM sets are constructed, and each of them corresponds to a different noise type (Subway, Babble, Car, Exhibition) and SNR value (0-30 dB in 2 dB intervals). A single HMM set which is closest to the test noisy speech is selected for recognition based on the estimated SNR value and

noise type of the test speech. Since MTR method combines a number of noise conditions to train a single HMM set, it tends to reduce the phonetic sharpness of the acoustic models in their probability density functions of the HMM. MMSR method can overcome the weakness of MTR by choosing a specific single HMM set which is most appropriate to the test noisy speech. However, the errors in selecting the closest HMM set will incur misrecognition, causing performance degradation in the MMSR.

3. Feature Compensation in the Presence of Channel Noise

Unlike the pervious study [11] where only additive noise is assumed to exist, we derive, in this study, a speech feature compensation method assuming that there is also channel noise in the test speech. The relation between training and test noisy speech is first derived in log-spectrum domain. Since the relation is non-linear, it is approximated using the VTS to obtain the mean vectors and covariance matrices of the test noisy speech given the statistics of training noisy speech obtained during the training. The statistics of the test noisy speech are used to obtain MMSE estimation of the training noisy speech, which is used as a feature vector for recognition after Discrete Cosine Transform (DCT).

3.1. Relation between Test and Train Noisy Speech

Log-spectrum vector $\mathbf{x}$ of the clean speech and $\mathbf{y}$ of the noisy speech are usually assumed to be related as follows:

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{x} - \mathbf{h})) \tag{1}$$

where $\mathbf{n}$ and $\mathbf{h}$ are the log-spectrum vector of additive and convolution noise, respectively, and $\mathbf{i}$ is a unity vector. Based on Equation (1), the log-spectrum vector of the test noisy speech $\mathbf{y}$ and the training noisy speech $\mathbf{y}_{Tr}$ can be expressed as follows, assuming that there is no channel noise in the training noisy speech for the convenience of analysis:

$$\mathbf{y}_{\mathrm{Tr}} = \mathbf{x} + \mathbf{g}_0(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}}) \tag{2}$$

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) \tag{3}$$

$$\mathbf{g}_0(\mathbf{x}, \mathbf{n}_{Tr}) = \log(\mathbf{i} + \exp(\mathbf{n}_{\mathrm{Tr}} - \mathbf{x})) \tag{4}$$

$$\mathbf{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) = \log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{x} - \mathbf{h})) \tag{5}$$

where $\mathbf{n}$ and $\mathbf{n}_{Tr}$ represent the additive noise contained in the test and training noisy speech, respectively. $\mathbf{n}_{Tr}$ should be determined during training, and $\mathbf{n}$ is estimated using test noisy speech in recognition.

By combining Equations (2) and (3), the test noisy speech can be expressed in terms of the training noisy speech as follows:

$$\mathbf{y} = \mathbf{y}_{\mathrm{Tr}} + \mathbf{h} + \mathrm{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) - \mathbf{g}_0(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}}) \tag{6}$$

$$\begin{aligned} \left[\mathrm{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) - \mathrm{g}_0(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}})\right]_i &= \log\left(\frac{\left[\mathbf{i} + \exp(\mathbf{n} - \mathbf{x} - \mathbf{h})\right]_i}{\left[\mathbf{i} + \exp(\mathbf{n}_{\mathrm{Tr}} - \mathbf{x})\right]_i}\right) \\ &= \log\left(\frac{\left[\exp(\mathbf{x}) + \exp(\mathbf{n} - \mathbf{h})\right]_i}{\left[\exp(\mathbf{x}) + \exp(\mathbf{n}_{\mathrm{Tr}})\right]_i}\right) \end{aligned} \tag{7}$$

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

From Equations (2) and (4),

$$\mathbf{y}_{Tr} = \mathbf{x} + \log\left(\mathbf{i} + \exp(\mathbf{n}_{Tr} - \mathbf{x})\right) \tag{8}$$

Taking the exponential of both sides of Equation (8),

$$\exp(\mathbf{x}) = \exp(\mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr}) \tag{9}$$

Substituting (9) into (7),

$$\begin{aligned} &\left[\mathbf{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) - \mathrm{g}_0(\mathbf{x}, \mathbf{n}_{\mathrm{Tr}})\right]_i \\ &= \log\left(\frac{\left[\exp(\mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr}) + \exp(\mathbf{n} - \mathbf{h})\right]_i}{\left[\exp(\mathbf{y}_{Tr})\right]_i}\right) \\ &= \left[\log\left(\mathbf{i} + \exp(\mathbf{n} - \mathbf{h} - \mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr} - \mathbf{y}_{Tr})\right)\right]_i \\ &= \left[\mathrm{G}(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{h}, \mathbf{n}_{Tr}\right]_i \end{aligned} \tag{10}$$

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

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

3.2. Statistics of Test Noisy Speech

From Equation (11), the mean and covariance of the test noisy speech can be estimated. Equation (11) is expanded using a first-order VTS around the initial value $\mathbf{n}_0, \mathbf{h}_0$ of $\mathbf{n}, \mathbf{h}$ and the mean of the training noisy speech $\boldsymbol{\mu}_{\mathbf{y}_{Tr}} = E\{\mathbf{y}_{Tr}\}$ to obtain the following equation.

$$\begin{aligned} \mathbf{y} \approx \mathbf{y}_{Tr} + \mathbf{h} + G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr}) + \\ \nabla_{\mathbf{y}_{Tr}} G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{y}_{Tr} - \boldsymbol{\mu}_{y_{Tr}}) + \\ \nabla_{\mathbf{n}} G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{n} - \mathbf{n}_0) + \\ \nabla_{\mathbf{h}} G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{h} - \mathbf{h}_0) \end{aligned} \tag{12}$$

Using Eq. (12), the mean $\boldsymbol{\mu}_y$ and covariance $\boldsymbol{\Sigma}_y$ of the test noisy speech can be expressed from the mean $\boldsymbol{\mu}_{y_{Tr}}$ and covariance $\boldsymbol{\Sigma}_{y_{Tr}}$ of the training noisy speech as follows:

$$\begin{aligned} \boldsymbol{\mu}_y \approx \boldsymbol{\mu}_{y_{Tr}} + \mathbf{h} + G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr}) + \\ \nabla_{\mathbf{n}} G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{n} - \mathbf{n}_0) + \\ \nabla_{\mathbf{h}} G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{h} - \mathbf{h}_0) \end{aligned} \tag{13}$$

$$\boldsymbol{\Sigma}_y = \left(\mathbf{I} + \nabla_{\mathbf{y}_{Tr}} G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})\right) \boldsymbol{\Sigma}_{y_{Tr}} \cdot \cdot \left(\mathbf{I} + \nabla_{\mathbf{y}_{Tr}} G(\boldsymbol{\mu}_{y_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})\right)^T \quad (14)$$

3.3. Estimation of Noise Parameter

Assuming also that the log-spectrum vector $\mathbf{y}$ of the test noisy speech is a mixture of Gaussian distributions, the distribution of $\mathbf{y}$ as a function of unknown noise vector $\mathbf{n}$, $\mathbf{h}$ can be defined using Eqs. (13) and (14),

$$p(\mathbf{y} \mid \mathbf{n}, \mathbf{h}) = \sum_{m=1}^{M} p_m N(\boldsymbol{\mu}_{y,m}, \boldsymbol{\Sigma}_{y,m}) \quad (15)$$

where $N(\boldsymbol{\mu}_{y,m}, \boldsymbol{\Sigma}_{y,m})$ is the m-th Gaussian distribution with a mean vector $\boldsymbol{\mu}_{y,m}$ and covariance matrix $\boldsymbol{\Sigma}_{y,m}$. p_m is the mixture weight of the m-th component. Note that the mean vector $\boldsymbol{\mu}_{y,m}$ and covariance matrix $\boldsymbol{\Sigma}_{y,m}$ are themselves fully parameterized by the noise vectors $\mathbf{n}$ and $\mathbf{h}$, which are treated just as parameters, not random variables; only the noisy speech vectors were treated as random variables.

Given a sequence of log-spectrum vectors for the test noisy speech , the log-likelihood for the sequence is defined as follows using Equation (15):

$$L(\mathbf{Y} \mid \mathbf{n}, \mathbf{h}) = \sum_{t=1}^{T} \log\, p(\mathbf{y}_t \mid \mathbf{n}, \mathbf{h}) \quad (16)$$

An iterative Expectation Maximization (EM) algorithm is used to re-estimate the noise vector maximizing Equation (16). In the EM algorithm, an auxiliary function $Q(\boldsymbol{\varphi}, \overline{\boldsymbol{\varphi}})$ is written as follows:

$$Q(\boldsymbol{\varphi}, \overline{\boldsymbol{\varphi}}) = E\{L(\mathbf{Y} \mid \overline{\boldsymbol{\varphi}}) \mid \mathbf{Y}, \boldsymbol{\varphi}\} = \sum_{t=1}^{T} \sum_{m=1}^{M} p(m \mid \mathbf{y}_t, \mathbf{n}, \mathbf{h}) \log p(\mathbf{y}_t, m \mid \overline{\mathbf{n}}, \overline{\mathbf{h}}). \quad (17)$$

The symbol φ represents the noise vector $\mathbf{n},\mathbf{h}$, which is already known and $\overline{\varphi}$ is the unknown noise vector $\overline{\mathbf{n}}, \overline{\mathbf{h}}$, which should be estimated. To re-estimate $\mathbf{n},\mathbf{h}$ in Eq. (17), the derivative of the auxiliary function with respect to $\overline{\mathbf{n}}, \overline{\mathbf{h}}$ must be taken and set equal to 0.

The noise vector $\overline{\mathbf{n}}, \overline{\mathbf{h}}$ derived is substituted into $\mathbf{n},\mathbf{h}$ in Equations (13) and (14) to adapt the mean and covariance of the test noisy speech. The likelihood function from Equation (16) and the auxiliary function from Equation (17) are consequently updated. This process is iterated until the log-likelihood function from Equation (16) converges. After the convergence, an MMSE estimation of the training noisy speech is performed and used for recognition.

3.4. MMSE of Training Noisy Speech

The MMSE of training speech $\mathbf{y}_{Tr}$ given the test speech $\mathbf{y}$ is expressed as follows:

$$\hat{\mathbf{y}}_{Tr,MMSE} = E(\mathbf{y}_{Tr} \mid \mathbf{y}) = \int \mathbf{y}_{Tr}\, p(\mathbf{y}_{Tr} \mid \mathbf{y}) d\mathbf{y}_{Tr} \quad (18)$$

From Equation (11),

$$\mathbf{y}_{Tr} = \mathbf{y} - \mathbf{h} - G(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{h}, \mathbf{n}_{Tr}) \quad (19)$$

Substituting Equation (19) into (18) and approximating $G(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{h}, \mathbf{n}_{Tr})$ by a VTS of order zero around $\boldsymbol{\mu}_{y_{Tr},m}$, the following relationship is obtained:

$$\hat{\mathbf{y}}_{Tr,MMSE} \cong \mathbf{y} - \mathbf{h} - \sum_{m=1}^{M} p(m \mid \mathbf{y}) G(\boldsymbol{\mu}_{Tr,m}, \mathbf{n}, \mathbf{h}, \mathbf{n}_{Tr}) \quad (20)$$

The DCT of the log-spectrum vector $\hat{\mathbf{y}}_{Tr,MMSE}$ is taken to find a 13th-order cepstrum vector. The c0 component in the cepstrum vector is replaced with log-energy. The delta and acceleration (delta-delta) coefficients of the cepstrum vector are also calculated to obtain a 39-dimentional feature vector which is used for the speech recognition experiments described in the next section.

4. Experimental Results

To verify the effectiveness of the proposed feature compensation method, experiments were conducted on the Aurora 2 database. For the feature vector, a noise-robust version of Mel-Frequency Cepstral Coefficients (MFCCs) called AFE (Advanced Front-End) was used. AFE is known to significantly reduce the word error rates in noisy speech recognition [12]. The 12th-order MFCCs with the 0th-order cepstral coefficient set aside are appended with the log-energy to form a 13th order basic feature vector along with their delta and acceleration coefficients to construct a 39th-order feature vector for each frame.

The acoustic models were trained using both the Complex Back End (CBE) and Simple Back End (SBE) scripts, which are each separately defined for the Aurora 2 database. For the SBE model, the HMM for each digit consists of 16 states with 3 Gaussian mixtures in each state. In addition, a three-state silence model with 6 Gaussian mixtures per state and a one-state pause model tied with the center state of the silence model are used. For the CBE, the number of mixtures in each state is increased to 20 and 36 for the digit and silence models, respectively. The hidden Markov model toolkit (HTK) was employed to train and test the HMM used in this study [13].

Table 1 WERs (%) of the proposed feature compensation methods using SBE models compared to conventional methods for Aurora 2 database.

Method	Set A	Set B	Set C	Ave.
Baseline	12.25	12.90	14.56	12.97
MTR	7.70	8.23	9.26	8.22
MMSR	6.78	9.56	8.17	8.17
MTR-MMSE (additive noise only)	7.54	7.75	9.18	7.95
MTR-MMSE+H (additive +channel noise)	7.61	7.52	8.82	7.81
MMSR-MMSE (additive noise only)	6.71	8.98	7.92	7.86
MMSR-MMSE+H (additive +channel noise)	6.42	8.24	7.45	7.35

Table 1 shows the word error rates (WERs) of the proposed feature compensation method in comparison with the conventional methods for the Aurora 2 database. MTR-MMSE/MMSR-MMSE are methods in our previous study [11], where only additive noise is adapted for the speech feature compensation while channel noise is additionally compensated in the proposed MTR-MMSE+H/MMSR-MMSE+H. MTR-MMSE/MTR-MMSE+H and MMSR-MMSE/MMSR-MMSE+H differ in the type of noisy speech HMM used for compensation.

Conventional MTR and MMSR method improve the performance of the baseline system which was trained using clean speech data. The baseline system scores 12.97% WER on average, whereas MTR and MMSR achieve WERs of 8.22 % and 8.17%, respectively. Although MMSR performs slightly better than MTR, their difference is minor.

By using the feature compensation, the performance of MTR and MMSR could be improved further. As shown in Table 1, MTR-MMSE and MMSR-MMSE achieve 7.95% and 7.86% average WERs by adapting the additive noise, providing 3.3% and 3.8% relative improvement over MTR and MMSR, respectively. By additionally compensating the channel noise, we could further improve the recognition performance of the MTR-MMSE/MMSR-MMSE. As shown in the table, MTR-MMSE+H achieve 7.81% average WER further reducing the WER of MTR-MMSE. This is mainly due to the performance improvement in Set B and C. The unexpected improvement in Set B seems to come from the initial value $\mathbf{h}_0$ of the channel noise which is defined as the difference between the test and training noisy speech. This initialization may have contributed to compensate the noise difference in Set B. Similar results have been also found in MMSR-MMSE+H as shown in Table 1.

The proposed methods were also applied to the noisy speech HMM trained with the CBE script to verify whether the proposed method could work well when the acoustic modeling becomes more complex. In Table 2, we can observe significant performance improvement when using the CBE script compared with the SBE script and it is more prominent in MTR than MMSR. The increased number of mixtures in each state of the HMM may have greatly contributed to sharpening the acoustic modeling in MTR. Although MMSR had comparable performance with MTR in the SBE script, MTR significantly outperforms MMSR in the CBE script.

Table 2. *WERs (%) of the proposed feature compensation methods using CBE models compared to conventional methods for Aurora 2 database.*

Method	Set A	Set B	Set C	Ave.
Baseline	11.58	12.10	13.68	12.20
MTR	6.04	6.82	7.22	6.59
MMSR	6.17	9.0	7.97	7.66
MTR-MMSE (additive noise only)	5.9	6.33	7.11	6.31
MTR-MMSE+H (additive+ channel noise)	5.92	6.23	6.69	6.19
MMSR-MMSE (additive noise only)	5.86	8.17	7.48	7.10
MMSR-MMSE+H (additive+ channel noise)	5.68	7.50	7.15	6.70

As in the SBE script, the performance of MTR and MMSR with the CBE script could be further improved by compensating the additive noise. As shown in Table 2, MTR-MMSE achieves 6.31% WER, providing 4.24% relative improvement over MTR. Similarly, MMSR_MMSE also show improved performance over MMSR. Since MTR outperforms MMSR in the CBE script, the WER of MTR-MMSE (6.31%) is shown to be much smaller than MMSR-MMSE (7.10%). We can also obtain additional performance gain by applying the channel compensation in the CBE script.

5. Conclusions

In this study, a VTS-based feature compensation method has been applied to noisy speech HMMs. In particular, we propose to adapt the channel noise for the compensation to improve the performance the previous method which takes into consideration only additive noise. The channel and additive noise were adapted to reduce the mismatch between the test noisy speech and the noisy speech HMM. The experimental results confirmed that the proposed feature compensation method is very effective in reducing the mismatch occurring in noisy speech recognition using MTR and MMSR based noisy speech HMMs. The feature compensation algorithm was applied to HMMs trained with the CBE script as well as the SBE script to test the robustness of the proposed method against varying HMM complexities and improved performance was found in both of them.

6. 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).

7. References

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 1, pp.113-120, 1979.

[2] M.J.F. Gales, "Model based techniques for noise-robust speech recognition," Ph.D. dissertation, University of Cambridge, 1996.

[3] P.J. Moreno, "Speech Recognition in noisy environments," Ph.D. dissertation, Carnegie Mellon University, 1996.

[4] D.Y. Kim, C.K. Un, N.S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication* vol. 24, no. 1, pp. 39–49, 1998.

[5] M. Akbacak, J.H.L. Hansen, "Environmental sniffing: noise knowledge estimation for robust speech systems," in *Proceedings of the International Conference of Acoustics, Speech and Signal Processing*, Hongkong, China, pp. 113-116, 2003.

[6] M. Akbacak, J.H.L. Hansen, "Environmental sniffing: noise knowledge estimation for robust speech systems," *IEEE Trans. Audio, Speech and Language Process.* vol. 15, no. 2, pp. 465-477, 2007.

[7] R.P. Lippmann, E.A. Martin and D.B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*., Dallas, Texas, USA, pp.705-708, 1987.

[8] H.G. Hirsch, D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the International Conference on Spoken Language Processing*, Bejing, China, pp.18-20, 2000.

[9] H. Xu, Z.H. Tan, P. Dalsgaard, B. Lindberg, "Robust speech recognition on noise and SNR classification – a multiple-model framework," in *Proceedings of INTERSPEECH,* Lisboa, Portugal, pp.977-980, 2005.

[10] H. Xu, X.H. Tan, P. Dalsgaard and B. Lindberg, "Noise condition dependent training based on noise classification and SNR estimation," *IEEE Trans. Audio, Speech, Language Process.,* vol. 15, no. 8, pp.2431-2443, 2007.

[11] Y. Chung and J.H.L. Hansen, "Compensation of SNR and noise type mismatch using an environmental sniffing based speech recognition solution," *EURASIP Journal on Audio, Speech, and Music Processing,* pp. 1-14, 2013.

[12] 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.

[13] S. Young, HTK: Hidden Markov Model Toolkit V3.4.1. Cambridge Univ. Eng. Dept. Speech Group, 1993.