MULTI-CONDITION TRAINING AND ADAPTATION FOR NOISE ROBUST SPEECH RECOGNITION

Ivan Kraljevski[1], Frank Duckhorn[1], Matthias Wolff[2] and Rüdiger Hoffmann[1]

[1]TU Dresden, Institute of Acoustics and Speech Communication, Dresden, Germany
[2]BTU Cottbus, Electronics and Information Technology Institute, Cottbus, Germany
{ivan.kraljevski, frank.duckhorn, ruediger.hoffmann}@tu-dresden.de
matthias.wolff@tu-cottbus.de

Abstract: In this paper, we investigated the recognition performance on speech distorted by noise with unknown characteristics, as found in many real-world situations. Speech recognition performance is notably degraded when used in conditions with mismatched test and training speech data. Various methods are proposed to overcome this problem for wide range of speech, speaker, channel and environmental conditions. The investigations presented in this paper, uses the UASR (Unified Automatic Speech Recognition and Synthesis) system to create and compare acoustic models regarding the noise robustness: model trained on clean speech data, model trained with multi-condition (M-C) noisy data and clean model adapted on (M-C) noisy speech data. The noise robustness of the models was investigated by phoneme recognition on speech data with added noise of certain type and SNR levels as well as noise of unseen characteristics. It was shown, as it was expected, that there is significant recognition performance degradation for the clean speech model, while the M-C trained model achieved the best possible recognition performance compared to others. It was observed also, that the adapted model could be successfully used for noisy speech recognition without the need of large amount of adaptation data.

1 Introduction

Speech recognition performance is notably degraded when used in conditions with mismatched test and training speech data. Various methods are proposed to overcome this issue for wide range of speech, speaker, channel and particularly environmental noise conditions. Generally, two different groups of noise robustness algorithms have been developed: noise filtering or speech enhancement and noise or environment compensation [1].

In the first group, after the standard feature analysis in the ASR's front-end (like PLP or MFCC [2]), various noise suppression methods could be used: Spectral Subtraction (SS) [3], Wiener filtering [4] and Minimum Mean Square Error Estimation (MMSE) [5], each of them assuming the availability of *a priori* knowledge of the noise spectral characteristics.

In the second group, noise compensation methods could be used in the acoustic modeling phase. Multi-condition (M-C) training is used to compensate the mismatched testing and training conditions and improve the recognition accuracy in acoustic noisy environments through the use of representative training data regarding the noisy conditions [6]-[9]. When no information about the noise characteristics is available, then, artificial noise with various characteristics and different values of SNR might be used for M-C training [10].

Model adaptation [11] is used when small amount of multi-condition speech data is available that cannot be used for complete model training. Parallel model combination (PMC) [12] creates noisy acoustic model from existing clean model by incorporating a statistical noise model. SPLICE (Stereo based Piecewise Linear Compensation for Environments) [13] produces an estimate of the corruption characteristics given the observed distorted speech cepstrum, while assuming the existence of stereo training data. Typically, the acoustic models are trained

firstly on clean speech data to provide high quality models that could be used later for adaptation. On the other hand, multi-condition training uses additive-noise contaminated speech training data to provide coarse compensation for the training and test data mismatch.

Conventional adaptation methods like Maximum *a posteriori* (MAP) [14] or Maximum Likelihood Linear Regression (MLLR) [15] are commonly used separately or in combination with noisy adaptation data. The MLLR algorithm performs good when smaller amount or sparsely populated adaptation data is available, while MAP requires larger amount of data for the same performance level. The used MAP algorithm updates the trained acoustic model parameters by joining the old with the new statistics parameters derived from the adaptation data.

Training or adaptation with specific noise type and level significantly improves recognition performance when the recognized speech is affected by similar conditions, but it degrades when training and recognition noise types do not match. If the training is performed with a variety noise types and levels, the robustness and the performance are both improved. The procedure is simple and most effective, but very time consuming and requires large amount of carefully collected and prepared training data.

In this paper, speech recognition in noisy environments was investigated in case when the accurate estimation of the noise type and characteristics is not possible. An UASR (Unified Automatic Speech Recognition and Synthesis) [16] system was used for model training and adaptation. The created acoustic models were compared regarding the noise robustness: model trained on clean speech data, model trained with M-C noisy data and clean model adapted on M-C noisy speech data. The robustness of all three models was investigated by phoneme recognition of speech data with added noise of certain type and level as well as noise of unseen characteristics.

The remainder of the paper is organized as follows, in Section 2 description of the experimental framework including the used noise type's characteristics and the adaptation algorithm is given. Next section presents the ASR system, acoustic modeling, the used speech databases and the results of the experiments carried for speech recognition. The last section summarizes the overall performance and concludes the results along with future directions for improvement.

2 Noise influence and model adaptation

2.1 Effects of additive noise

Additive noise affects the speech recognition due to mismatch between training and recognition speech data in the feature representation domain and also to its randomness which cause speech information loss. Speech recognition performance degrades, because the clean acoustic models do not model the noisy speech accurately. The majority of the noise robust speech recognition methods are focused on reducing this mismatch. Also, the information loss caused by noise introduces degradation even in the case of optimal mismatch compensation. Detailed analysis is given in [17] about the additive noise effects on the feature vectors parametrization and the probability distributions.

It is assumed, that the speech and noise signals are uncorrelated, that they could be linearly combined in spectral domain and that the convolution noise is not considered. The output energy of the filter i in the filter bank at frame t, corresponding to the noisy speech $Y_i(t)$ can be written as a function of the energy of the clean speech $S_i(n)$ and the noise $N_i(t)$:

$$Y_i(t) = S_i(t) + N_i(t) \tag{1}$$

and the relation in the log domain $x_i = \log(X_i)$ is described by the equation:

$$y_i(t) = \log[\exp(x_i(t)) + \exp(n_i(t))] \quad (2)$$

Therefore, the effect of the additive noise consists of a nonlinear transformation of the representation space in the log domain which produces a mismatch between the clean and the noisy conditions. Generally, in the feature extraction domain the additive noise:

- produces a non-linear distortion of the representation space;
- masks the speech signal in the regions where noise level is greater than speech level and the log-energy of the noisy speech is similar to that of the noise;
- slightly affects the noisy speech where the speech level is greater than noise level;
- affects static features (especially energy) more than dynamic features.

and the additive noise effects on probability distributions:

- causes a displacement of the mean values;
- the standard deviation is reduced due the non-uniform compression caused by noise;
- the noisy pdf is distorted and it is not a Gaussian distribution due to the non-linear effect of the noise.

In order to see the separate contribution of the effects, information loss and the model mismatch, speech recognition experiments should be performed on clean acoustic models (baseline) and M-C retrained model. The degradation for the phoneme recognition performance on M-C models is related to the information loss caused by noise, while the recognition performance degradation on clean acoustic model represent the degradation due to both, the clean and noisy condition mismatch and the information loss.

2.2 Model adaptation

The MAP adaptation method [14] updates the HMM model parameters by joining known information (the old parameters) with the statistics derived from the adaptation data. The adaptation process consists of two stages. In the first step, the statistics required for computing the distribution weight, mean and covariance are gathered. The mean value $\tilde{\mu}_j$ (3) and covariance matrix elements $\tilde{\Sigma}_j(x,y)$ (4) of a Gaussian distribution j, over N_j adaptation data samples $o_{n,j}$, are calculated from the following statistics:

$$\tilde{\mu}_j = \frac{1}{N_j} \cdot \sum_{n=1}^{N_j} o_{n,j} \quad (3)$$

$$\tilde{\Sigma}_i(x,y) = \frac{1}{N_j - 1} \left(\sum_{n=1}^{N_j} o_{n,j}(x) \cdot o_{n,j}(y) - \frac{\sum_{n=1}^{N_j} o_{n,j}(x) \cdot \sum_{n=1}^{N_j} o_{n,j}(y)}{N_j} \right) \quad (4)$$

In the second step the statistics from adaptation data are combined with the old statistics from the HMM model using data-dependent weighting coefficient. The data dependency is designed to weight the statistics with higher population toward new parameters and with lower population toward the original parameters. The new mean (5) and covariance (6) for the distribution j presents weighted sum of the old and the new statistics:

$$\hat{\mu}_j = \frac{n_j}{n_j + \rho_\mu} \tilde{\mu}_j + \frac{\rho_\mu}{n_j + \rho_\mu} \mu_j \quad (5)$$

$$\hat{\Sigma}_j = \frac{n_j}{n_j + \rho_\Sigma} \tilde{\Sigma}_j + \frac{\rho_\Sigma}{n_j + \rho_\Sigma} \Sigma_j \quad (6)$$

The data-dependency of the weighting coefficients is realized by the relevance factor ρ_μ and ρ_Σ. Theirs values mark the points where the data count of the adaptation data has the same weight as the old parameter. Higher values of ρ give more weight to the prior information, the old parameters. The main problem with the MAP adaptation method is that it is an unconstrained method and updates therefore only those parameters where observations exist. It requires a relatively large amount of adaptation data in order to be effective for sparsely occupied Gaussian distributions.

3 Experimental setup

UASR is a speech dialogue system with synthesis and recognition components that uses common databases. The system uses arc-emission HMM with one single Gaussian density per arc and an arbitrary topology. The structure is built iteratively during the training process by state splitting from an initial HMM model [16]. The advantage of these structures as a underlying concept lies in their enhanced capability of modeling trajectories in the feature space.

The following system setup was used for the experiments. The clean and noisy speech signals, sampled at rate of 16 KHz and 16 bits per sample, were divided in 32 ms wide frames with a frame period of 10 ms and processed with a Hamming window. The band from 300 to 8000 Hz was covered with 31 Mel DFT filters and at the output of each channel the log of the energy was computed. The obtained feature vectors and theirs delta values were standardized to a mean of zero and standard deviation of one, giving a final feature vector with dimension of 60. Principal Component Analysis (PCA), as an orthonormal transformation that provides a linear mapping for dimension reduction and de-correlation, was used to bring the number of components to 24. The effectiveness of PCA in pattern recognition lies in its ability to decorrelate feature parameters and relegate most of the random structures (noise) to trailing components while extracting systematic patterns to leading ones. The acoustical model consists of 42 monophonic models, one pause model and one garbage model. The structure is built iteratively during the training process on clean and noisy speech by state splitting from an initial HMM model in 2 splits, therefore giving 12 Gaussian distributions per phoneme or in total 516.

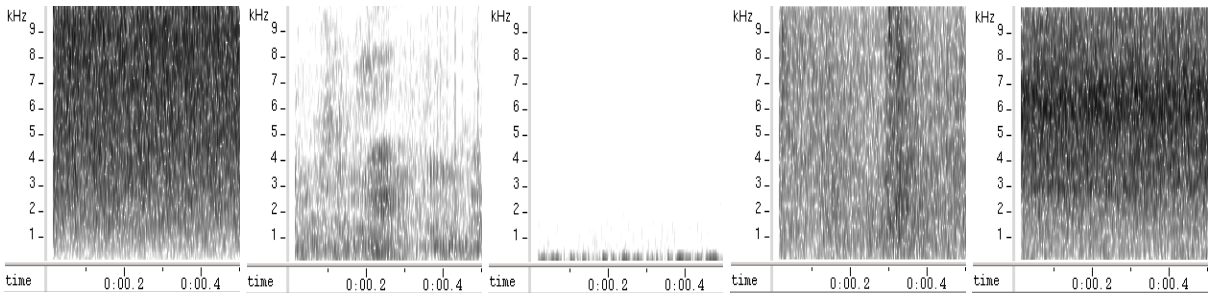
3.1 Databases and acoustic modeling

For the acoustic models training and evaluation, small data subset of 1537 turns with total duration of around 3 hours 34 min from the CD 2.0 of Verbmobil German Database [18] was used. The test set consists of 70 sentences (with 577 seconds) and the training set of 1396 sentences (with 11624 seconds). The development set with 71 sentences (659 seconds) was contaminated with M-C additive noise (4 sentences per noise type and level). Firstly, the training set with clean speech was used to produce baseline clean acoustic model with two splits from the initial HMM model. Then, the same speech database contaminated by additive noise with equal type and SNR level contribution (~70 sentences per type and level) was used for the M-C model training in the same way as for the clean model. Afterward, the clean model was adapted using MAP algorithm on the M-C development set. The performances of the models were evaluated and compared in phone recognition experiments, because this is of fundamental importance as they reveals the quality of the acoustic modeling.

3.2 Noise types used for M-C training and adaptation data

For the purpose of speech recognition on baseline (clean), retrained and adapted acoustic models, multi-condition training data set was prepared. It consists noisy speech samples of four different spectral characteristics ("babble", "factory", "volvo", "white") and clean speech with four SNR levels (5, 10, 15 and 20 dB), taken from the NOISEX-92 database [19].

A non-stationary noise example ("buccaneer") is also used in the experiments as unknown noisy environment condition included neither in the training nor adaptation process. The different noise type's spectral characteristics are presented on Figure 1. It could be noticed, that in all cases the additive noise covers the whole available bandwidth except in the case of noise type "volvo". Here, significant noise signal energy is spread in narrow part of the lower frequency band, where only small part of the speech bandwidth is distorted. The added "volvo" noise signal distorts the speech waveform significantly due to the large signal amplitude, but only for lower frequencies, therefore the phoneme recognition will be lesser affected compared with other noise types.



a. white b. babble c. volvo d. factory e. buccaneer

Figure 1 - Spectrograms of the noise used for M-C data creation (a-d) and unseen noise (e)

In order to precisely combine the additive noise with the clean speech signal, the gain parameter g was used to compensate the difference in the signal energy levels. To compute the value of g, desired SNR value of the noisy signal could be used as in (7).

$$SNR = 10\log_{10}\left(\frac{g^2 \cdot P_s}{P_n}\right) \quad \text{giving} \quad g = 10^{\frac{SNR}{20}} \cdot \sqrt{\frac{P_n}{Ps}} \quad (7) \quad \text{or} \quad g \simeq \frac{P_{sn} - P_n}{P_s} \quad (8)$$

In (8), another formula to estimate the gain parameter is shown, where, P_{sn} is the power of noisy signal, P_s and P_n are the power of the clean and noise signals, respectively. Therefore, the M-C data set consists of 5 different noise types with 4 SNR levels, giving 20 different additive noise environmental conditions.

4 Experiments on UASR

The UASR system was used for phoneme recognition performance evaluation in order to show the effectiveness of the acoustic features modeling. The evaluation parameter accuracy of the recognized label (phoneme) sequence (LSA) was calculated by the number of all phonemes in the reference sequence N^{all}, removed phonemes N^{del}, substituted phonemes N^{sub} and the number of inserted phonemes N^{Ins}. These numbers are calculated with sequence alignment using Levenshtein distance (9):

$$LSA(\%) = 1 - \frac{(N^{del} + N^{ins} + N^{\text{sub}})}{N^{all}} \cdot 100 \quad (9)$$

On Table 1 the phoneme recognition results are presented. First, the small test set of VM2 database consisted of 70 clean speech sentences with total duration of 577 seconds was used to evaluate the baseline performance of clean, M-C and adapted acoustic model. Then, test data contaminated with each noise type ("babble", "factory", "volvo", "white") and level (5, 10, 15 and 20 dB) was evaluated and the results were averaged in order to make comparison between all three used acoustic models.

Average (20-5 dB)	Clean	White	Babble	Volvo	Factory	Buccaneer	Average
Clean model	46,10	19,65	20,18	44,13	26,63	22,63	26,64
M-C model	43,10	27,45	31,90	42,88	32,48	27,85	32,51
Adapted model	40,30	24,90	32,43	42,38	32,20	24,68	31,32

Table 1 - Recognition results for Label Sequence Accuracy (%)

From Table 1 it could be seen that, the M-C trained acoustic model provides largest phoneme recognition accuracy over all seen and especially for the unseen noise conditions. The 95% confidence interval is in the range between ± 1,1 % and ± 2,9 % for all results presented in the table. On Figures 2, 3 and 4 the evaluation results are presented for each acoustic model with included noise samples ("buccaneer") unseen in the training process. Generally, it could be noticed that "volvo" noise has smallest influence on the recognition performance of all evaluated acoustic models. Because of its spectral characteristics, only small portion of the bandwidth is corrupted and thus few feature vector components are affected. As seen in Figure 2 it is clear that the recognition performance for the other conditions decreases regarding the SNR level, with "babble" noise recognition performance decreasing steeper for lower SNR values.

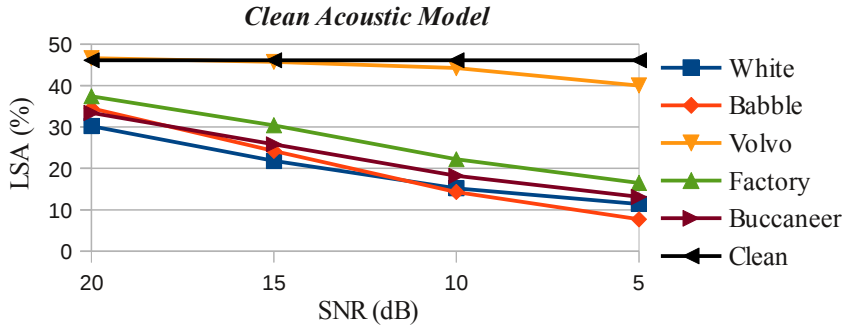


Figure 2 - Phoneme recognition performance on Clean Acoustic Model

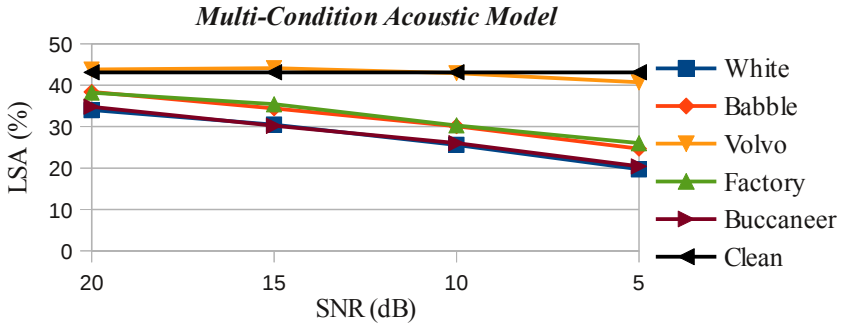


Figure 3 - Phoneme recognition performance on M-C Acoustic Model

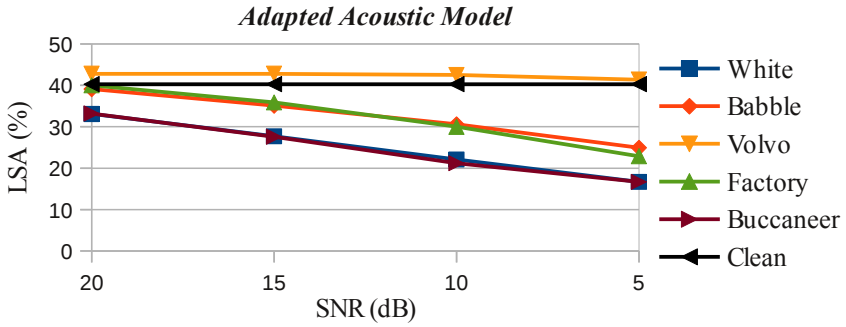


Figure 4 - Phoneme recognition performance on Adapted Acoustic Model

It is shown on Figure 3 for the M-C acoustic model, that in the case of "babble" and "factory" the recognition accuracy is more improved compared against "white" and "buccaneer" noise conditions. The reason is that the model is better trained on the specific spectral characteristics of these noise conditions. While for the "volvo" condition there is a slight performance decrease, almost equaling the results for clean speech recognition on M-C acoustic model.

The experimental results for the adapted acoustic model presented on Figure 4, shows that the difference in accuracy improvement for "babble" and "factory" is even more notable and here the performance results for "volvo" also surpasses the results on clean speech. This proves that using small amount of development speech data contaminated with M-C noise to adapt clean speech acoustic model could improve the recognition performance for all, particularly for those noise conditions with distinctive spectral characteristics. In all cases, the unseen noise condition "buccaneer" produces similar results as "white" noise condition, due to their spectral characteristics similarities.

SNR	LSA (%)			Relative improvement (%)	
	Clean	M-C	Adapted	M-C	Adapted
20 dB	36,42	37,84	37,64	3,899	3,350
15 dB	29,58	34,92	33,82	18,053	14,334
10 dB	22,82	30,98	29,28	35,758	28,309
5 dB	17,74	26,3	24,52	48,253	38,219

Table 2 - Relative performance improvement (%) of M-C and adapted model

Table 2 presents the relative improvement of the M-C and adapted model regarding the recognition performance on clean model. It is obvious that M-C model provides larger improvement compared to the adapted model, despite the difference is small. But, the adapted model was created using very limited amount of speech data in significantly shorter time. The 95% confidence interval is in the range between ± 1.1 % and ± 2.9 % for all results presented in the table.

5 Conclusions

The speech recognition performances were evaluated over various noise types and SNR levels as well as for unseen noise characteristics. It was shown, as it was expected, that there is a significant degradation of the recognition performance in the case of acoustic model trained on clean speech. The multi-condition trained acoustic model, achieved the best possible recognition performance as compared to the other two models: clean and adapted. Also, it was shown that the adapted model could be also successfully applied with slightly lower recognition accuracy, but without the need of large amount of adaptation data. Further improvements of the phoneme accuracy rate could be achieved by using additional speech enhancement methods combined with the noise compensation algorithms. The overall performance for unknown noisy conditions could be improved with careful preparation of the M-C data set for training or adaptation where it will consists of as many as possible noisy speech samples and SNR levels derived from real noisy environments.

References

[1] J. Rajnoha, "Multi-Condition Training for Unknown Environment Adaptation in Robust ASR Under Real Conditions" in *Acta Polytechnica,* 2009, Vol. 49 No. 2–3/2009. p. 3-7.

[2] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech" in *J. Acoust. Soc. Am.*, vol. 87, no. 4, p. 1738–1752, 1990

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[4] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouver, H. Kelleher, D. Pearce, and F. Saadoun, “Evaluation of a Noise-robust DSR Front-end on Aurora Databases,” in *Proc. ICSLP’2002*, Denver, CO, 2002, pp. 17–20.
[5] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator" in *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-32 (1984), no. 6, December 1984.
[6] R. Lippmann, E. Martin, and D. Paul, "Multi-style Training for Robust Isolated-word Speech Recognition", in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Apr. 1987, vol. 12, pp. 705–708.
[7] D. Pearce and H.-G. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions," in *Proceedings of the 6th International Conference on Spoken Language Processing*, pp. 29-32, 2000.
[8] X. Cui and Y. Gong, "Variable Parameter Gaussian Mixture Hidden Markov Modeling for Speech Recognition", in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, 2003
[9] J. Ming, P. Jancovic, P. Hanna and D. Stewart, "Modeling the Mixtures of Known Noise and Unknown Unexpected Noise for Robust Speech Recognition" *in Proc. European Conference on Speech Communication and Technology (Eurospeech’2001)*, Aalborg, Denmark, September 2001, p. 579–582.
[10] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust Speaker Recognition in Noisy Conditions", *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1711–1723, Jul. 2007.
[11] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation Within the MLLR Framework", *in Comput. Speech Lang.,* vol. 10, pp. 249–264, 1996.
[12] M. J. F. Gales and S. J. Young, "Parallel Model Combination for Speech Recognition in Noise" in *Technical report CUED/F-INFENG/TR 135*, Cambridge, England, 1993.
[13] L. Deng, A. Acero, L. Jiang, J. Droppo, and X.-D. Hunag, "High-performance Robust Speech Recognition Using Stereo Training Data", *in Proc. ICASSP’2001*, Salt Lake City, UT, 2001, pp. 301–304.
[14] J. L. Gauvain, C-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", in *IEEE Transactions on Speech and Audio Processing*, Vol.2, No.2, April 1994, 291-298.
[15] C. J. Legetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMM’s", in *Computer, Speech and Language*, vol. 9, pp. 171-186, 1995.
[16] M. Eichner, M. Wolff, S. Ohnewald, and R. Hoffmann, "Speech Synthesis Using Stochastic Markov Graphs”, in *Proc. Int Conf. on Acoustics, Speech and Signal Processing (ICASSP 2001)*, Salt Lake City, UT, USA.
[17] A. de la Torre, J. C. Segura, C. Benitez, J. Ramirez, L. Garcia and A. J. Rubio, "Speech Recognition Under Noise Conditions: Compensation Methods, Robust Speech Recognition and Understanding", *in Book edited by: Michael Grimm and Kristian Kroschel*, ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007
[18] T. Bub, J. Schwinn, "VERBMOBIL: The Evolution of a Complex Large Speech-to-Speech Translation System", in *Proc. Int. Conf. on Spoken Language Processing*, 1996, Philadelphia, PA, USA, October 1996, vol. 4, pp. 2371-2374.
[19] A. Varga and H. Steeneken, "Assessment for Automatic Speech Recognition II: Noisex-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems", *in Speech Commun.*,vol. 12, no. 3, pp. 247–251, 1993.