Robust Speech Recognition Based on Noise and SNR Classification - a Multiple-Model Framework

Haitian Xu, Zheng-Hua Tan, Paul Dalsgaard, Børge Lindberg

Center for TeleInFrastructure (CTIF), SMC-Speech and Multimedia Communication,
Aalborg University, Denmark
{hx, zt, pd, bli}@kom.aau.dk

Abstract

This paper presents a multiple-model framework for noise-robust speech recognition. In this framework, multiple HMM model sets are trained - each identified by a noise type and a specific Signal-to-Noise Ratio (SNR) value. This, however, does not increase the computational complexity of the recognition process since only one model set is selected according to the noise classification and SNR estimation. The optimal number of model sets is first identified on the basis of the Aurora 2 database. With only three model sets for each noise type, the framework shows superior performance to Multi-style TRaining (MTR) when testing on known noise types but lower performance on unknown noise types. To overcome this drawback, a modified Jacobian method is proposed to adapt the selected HMM models to the test environment. Furthermore, given the fact that MTR often gives relatively stable performance for unknown noise types, a combined technique is applied in which interpolation between the MTR and the adapted models is performed. This combined technique gives more than 24% performance improvement as compared to MTR.

1. Introduction

The deployment of Automatic Speech Recognition (ASR) in mobile devices imposes a much more varying acoustical environment than other typical settings such as in desktop computers. The performance of ASR systems in general in these cases is highly influenced by various noise types that each spans a wide range of Signal-to-Noise Ratio (SNR) values, causing huge mismatches between the data used for training and under real-life use.

The goal of applying the Multi-style TRaining (MTR) method [1] is to recover the degraded ASR performance by training the acoustic models by using a speech corpus corrupted with acoustic noise of the types likely to be encountered during use. The MTR method in general improves the ASR performances for the trained (known) noise types as well as for untrained (unknown) noise types on the one hand, but on the other hand the HMM model sets are inevitably built with flatter Probability Density Functions (PDF) which reduce the discriminability among the speech models. One way to partly overcome this drawback is to sub-divide the entire noise space into several smaller clusters (or noise types), and then train an HMM model set for each cluster. With the aim of maintaining the computational efficiency, the Multiple-Model Framework (MMF) is often combined with a Noise Classification (NC) technique to select one model set to be used during recognition [2].

The research of this paper extends the above NC based MMF (NC-MMF) by additionally taking the SNR-range of the individual noise types into account, namely SNC-MMF. The architecture of the SNC-MMF technique is illustrated in Fig.1 where the HMM Model Database (HMD) contains a number of HMM model sets each trained on data corresponding to a known noise type and a chosen SNR value. The NC and the SNR estimator shown in the figure estimate the noise type and the SNR value of the noise contaminated input signal, respectively. An HMM model set is then selected from the HMD and used for recognition. With this more detailed partitioning of the training database the PDF's of each HMM set have less variance, and a better model discriminability is expected.

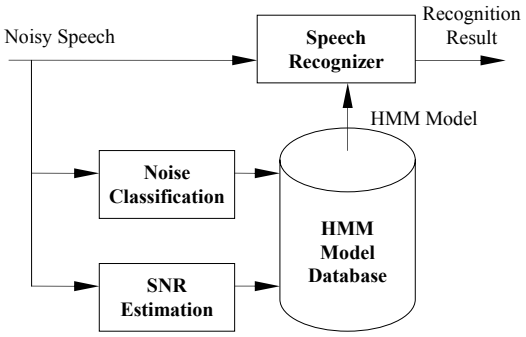

Fig. 1 Architecture of SNC-MMF

In this paper, a number of experiments are conducted with the aim of selecting a sufficient number of model sets for the model database, while still maintaining acceptable ASR performance.

The experiments verify that the SNC-MMF gives good ASR performance for known noise types but low for unknown noise types due to the noise emphasis given in its individual model sets. With the goal of counteracting this discrepancy, adaptation based on the Jacobian method [3] [4] is introduced in this paper, and a modified version – the Zero-noise-level difference Jacobian (Z-JAC) is proposed. For unknown noise types the MTR models generally give better ASR performance than the SNC-MMF models, directing us to suggest both MTR model interpolation and model adaptation in a combined framework.

The remainder of the paper is organized as follows. Section 2 provides an analysis of the SNC-MMF framework. Based on this Section 3 introduces adaptation based on the Jacobian method and the MTR model interpolation with the goal of improving the overall ASR performance regardless of noise type. The results from the experiments are given and discussed in Section 4 and the conclusion is given in section 5.

2. The SNC-MMF framework

2.1. SNC-MMF configuration

The SNC-MMF models are constructed on the basis of the Aurora 2 database [5] which consists of connected English digits corrupted with a number of artificially added noise types. The four noise types occurring in test Set A ("Subway", "Babble", "Car" and "Exhibition") are treated as known noise types in the following experiments.

In this work a simple Voice Activity Detection (VAD) based SNR estimator is used. NC is achieved by a cepstral GMM based noise classifier with four mixtures trained on the noise files in [5]. The noise is classified on the basis of the first 10 frames of the non-speech segments in each test utterance.

The confusion matrix of the noise classifier on test Set A is shown in Table 1. Only minor classification errors occur, except for the Car-noise (error in 3.62% of the cases).

Table 1: GMM noise classification results for the known noise types (test Set A)

Results / Test Noise	Subway	Babble	Car	Exhibition
Subway	99.78%	0	0	0.22%
Babble	0.01%	98.88%	0.96%	0.15%
Car	0.34%	3.62%	95.52%	0.52%
Exhibition	0	0	0	100%

2.2. Performance analysis for the known noise types

With the goal of investigating the performance of the SNC-MMF, a set of training and recognition experiments are conducted. The HTK speech recogniser [6] is used for these experiments applying the scripts provided by [5]. Each digit is modelled by 16 HMM states each with three Gaussian mixtures. The speech features are the normally used 39-dimensional MFCC vector.

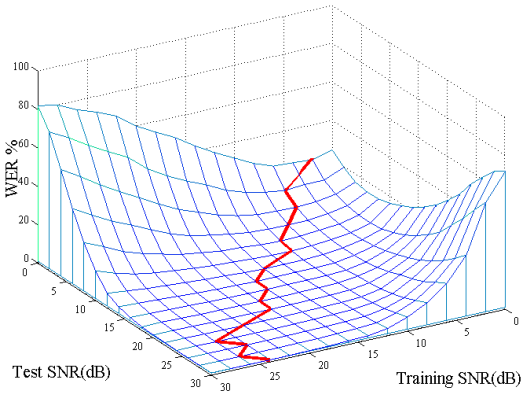

Fig.2 Word Error Rate (WER)-surface for Set A "Car" noise with different SNR combinations of training and test, and the lowest WER performance line

For each known noise type (one of the four occurring in Set A), noise data and clean speech training data (8440 files) are artificially added with SNR values ranging from 0dB to 30dB with 2dB intervals, resulting in an HMM model set for each of the SNR values. Recognition experiments are conducted with a separate speech corpus (1001 files) corrupted by the same type of noise and SNR range as used during training.

As an example, the Word Error Rate (WER) results for "Car" noise are shown in Fig.2. Similar WER-surfaces are obtained for the "Subway", "Babble" and "Exhibition" noise types.

From the results in Fig.2 it is observed that:

- No single model set can be chosen to give acceptable WER for all test SNR values. The lowest WER performance line on the WER-surface lies approximately along the diagonal from point (30, 30) to (0, 0) indicating that multiple model sets - each modelling a specific SNR value - are needed in order to achieve the lowest averaged WER for all the test SNR values.
- A relatively flat low-WER surface range exists around the lowest WER performance line indicating that the SNC-MMF framework is relatively insensitive to inaccurate SNR value estimations. Though not shown here, this is also observed for the other three known noise types.
- The lowest WER performance does not necessarily occur for an exact SNR value match between training and test data. This is further illustrated in Fig.3 where details from the experiments involving all four known noise types are given. The "Best performance Point" (BP) for each of the known noise types, defined as the point with the lowest WER performance for a given training SNR, is normally not the point where training and test SNR values are equal. It is noted that a majority of the BP's are above the diagonal reference line, especially when the SNR value is low. This finding is further exploited in Section 3 in conducting model adaptation.

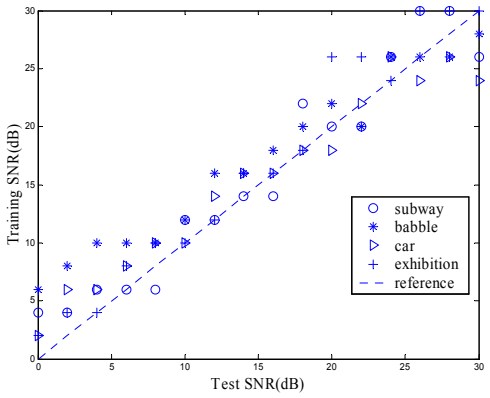

Fig.3 BPs for the four known noise types

Given that the SNC-MMF is relatively insensitive to inaccurate SNR estimation it is potentially feasible to reduce the number of model sets. With the aim of minimizing the number of models for each of the known noise types, an exhaustive search is carried out over the WER-surface for each chosen number of model sets (from one to sixteen), which identifies the combination of SNR model sets that results in the lowest WER.

The results are shown in Fig.4. A drastic decrease in WER is observed when the number of model sets starts increasing from one while there is almost no significant performance improvement observed when choosing more than five model sets. This indicates that 3-5 model sets are enough to achieve acceptable overall performance.

In the remainder of this paper, three model sets are used for each of the known noise types, which results in a relative 4% performance drop only, as compared to using all 16 models. An interesting fact observed from the search is that across all the four known noise types, the SNR values of the three models are similar and all close to 5dB, 10dB and 20dB.

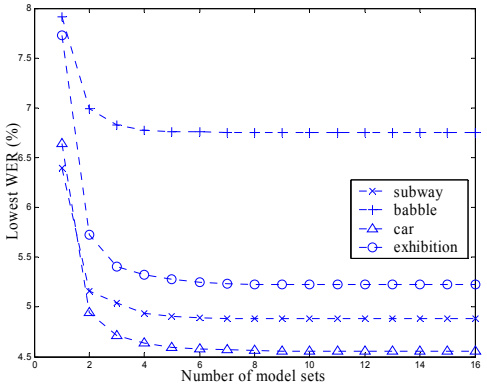

Fig. 4 Best performance with different number of HMM model sets for each noise type

3. Dealing with noise type mismatch

This section deals with noise type mismatch between the selected HMM model set and the testing environments. Mismatch occurs as a result of either noise classification errors or the fact that the noise type is unknown. Two methods, namely Jacobian adaptation and model interpolation, are introduced.

3.1. Jacobian adaptation

Jacobian adaptation (JAC) [3] [4] is a commonly used method aimed at adapting models to the varying acoustical environments. For simplicity, this research considers the JAC adaptation on the 13 static cepstral components in the mean vectors only.

Assume that the C_{N1} and C_{N2} with components $c_{i,x}$, $0 \le i \le 12$ and $x \in \{N1, N2\}$, are the averaged noise cepstral vectors of the known training noise $N1$ and the test noise $N2$ respectively, and that the corresponding mean vectors of the HMM model Gaussian mixtures are μ_{S+N1} and μ_{S+N2}. Then by linearly approximating the nonlinear cepstral distortion using a first-order Taylor series, the general JAC adapts the mean vector μ_{S+N2} as follows:

$$\mu_{S+N2} = \mu_{S+N1} + J_{N1} \times (C_{N2} - C_{N1}) \qquad (1)$$

where J_{N1} is the Jacobian matrix obtained during training. Generally JAC will result in significant approximation errors when the difference between C_{N1} and C_{N2} is large. However, this is less likely to occur in the present SNC-MMF approach due to the expected smaller environmental difference between training and test in a multiple model framework.

In the SNC-MMF, given that environmental SNR differences can be handled by using multiple model sets for different SNRs and that the exactly matched SNR training does not always guarantee the best performance, it is unnecessary and in some cases even harmful to eliminate them as JAC aims at in its standard form. Therefore we suggest a modified Jacobian adaptation (Z-JAC) which only attempts to eliminate the cepstral distortion caused by the mismatch in the noise types. Z-JAC assumes the same (noisy) speech energy in the two environments which otherwise can be achieved by energy normalization, and then simply sets the noise energy component $c_{0,x} = 0$ in Eq. (1) but leaves the other cepstral components unchanged. This has the effect of adapting the trained model set to an environment with the test noise type while the same SNR value as the trained model set.

3.2. Model interpolation

In [1] it has been observed that by mixing the information characterizing data collected in different noisy environments, the MTR models are generally robust to unknown types of noise. Interpolation is therefore introduced to bring further robustness to SNC-MMF models on the basis of the trained MTR models. The interpolation is defined as follows:

$$f_I(O) = \alpha f_N(O) + (1-\alpha) f_{MTR}(O) \qquad (2)$$

Given the observation O in Eq.(2), $f_I(O)$, $f_N(O)$ and $f_{MTR}(O)$ are respectively the PDF's for the finally interpolated model set, for the selected (or Z-JAC adapted) known noisy model set and for the MTR model set. The interpolation factor α should ideally be expressed by the correlation between the test and training noises. In this work, a fixed value of 0.4 is empirically chosen for simplicity.

4. Experiments

The SNC-MMF is evaluated on the Aurora 2 database. As described in section 2, the experimental settings for SNC-MMF are the same as the MTR in [5] that is therefore used here as the baseline. A set of recognition experiments are conducted for test Set A (including four known types of additive noise), Set B (including four unknown types of additive noise) and set C (including one known and one unknown type of noise with convolutional noises). Using the same weighting as in [5], the overall performance is calculated according to 0.4*(A+B) + 0.2*C.

The comparisons between the baseline MTR and a number of SNC-MMF settings for Set A are illustrated in Fig.5. The three settings for SNC-MMF models are: i) "SNC-MMF1" where SNC-MMF is given the a-priori knowledge of the known noise type and a given SNR value of the test data (i.e. no noise classification and SNR estimation errors), ii) "SNC-MMF2" where the known noise type is given but the SNR value is estimated by the SNR estimator, and iii) "SNC-MMF" where both the noise type and the SNR value are estimated. The results show that, a) the SNC-MMF offers significantly better performance than the MTR for the known

noise types, b) with only a minor deviation in performance among the three SNC-MMF settings, the SNC-MMF performance is robust to the provided SNR estimator and noise classifier, and c) the influence of the relatively small noise classification errors among the four types of noise - as given in Table 1 - can still be observed indicating that accurate noise classification is vital for this framework.

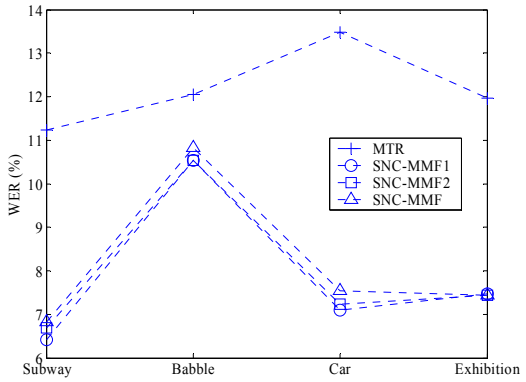

Fig. 5 Comparisons among different SNC-MMF settings and the MTR for the four known noise types in Set A

This sensitivity to training-testing noise type difference is further confirmed by the lower performance for the unknown noise types in Set B as shown in Table 2.

Table 2 WER (%) for different test sets and relative improvement (%) compared to MTR

Methods \ Sets	Set A	Set B	Set C	Average	Improv.
MTR	12.18	13.73	16.22	13.61	--
SNC-MMF	8.15	16.99	13.56	12.77	**6.2**
JAC	8.80	13.47	13.37	11.58	**14.9**
Z-JAC	8.35	13.01	12.94	11.12	**18.3**
Model Interp.	8.11	13.43	12.57	11.13	**18.2**
Model Interp. +Z-JAC	8.05	11.41	12.45	10.28	**24.5**

Table 2 compares the WER performance over a number of test sets for MTR, SNC-MMF, Jacobian adaptation and model interpolation. The basic SNC-MMF shows a significant improvement over MTR for Set A (the known noise types) but lower performance for Set B (the unknown noise types). Deploying JAC together with the SNC-MMF improves the performance for the unknown noise types and employing the Z-JAC adaptation further improves the performance slightly but steadily. Model interpolation shows improved performance for both known and unknown types of noises, and when combined with Z-JAC it gives a relative WER reduction of more than 24% as compared to the MTR.

5. Conclusions

This paper introduces the SNC-MMF framework to improve the noise robustness of speech recognition in general. During training, different HMM model sets are built for a number of combinations of noise types and SNR values. The computational complexity during the recognition process is, however, kept low by selecting only one model set on the basis of the estimation of noise type and SNR value in the test environments. The Aurora 2 database is used in finding the optimal number of model sets and the results consistently show that with only a limited number (from three to five) of model sets for each noise type, this framework is capable of achieving a good performance across a wide range of SNR values.

Using only three model sets leads to significant improvements for the known noise types as compared to the MTR while lower performance is observed for the unknown noise types. The introduction of Jacobian model adaptation (both JAC and Z-JAC) results in more robust models when dealing with unknown noise types. Finally, MTR model interpolation combined with Z-JAC is introduced with the goal of exploiting the robustness properties of MTR against unknown noise types. This latter combination in particular exhibits superior performance compared to standard MTR.

6. Acknowledgements

This work is supported by a PhD grant from the CNTK (Centre for Network and Service Convergence) project that is partly granted by the Danish Ministry of Science, Technology and Development, partly the participating industrial partners. Furthermore, the authors here are also grateful to Prof. John H. L. Hansen, University of Colorado at Boulder, for the valuable discussions.

7. References

[1] R. P. Lippmann, E. A. Martin, and D. B. Paul. *Multi-style training for robust isolated-word speech recognition*. Proc. ICASSP-87, pages 705--708, 1987

[2] M. Akbacak, J.H.L.Hansen. *Environmental sniffing: noise knowledge estimation for robust speech systems* Proceedings. (ICASSP '03). Volume 2, pages 113 -116, April 6-10, 2003

[3] S.Sagayama, Y.Yamaguchi and S.Takahashi. *Jacobian adaptation of noisy speech models* IEEE Workshop on Automatic Speech Recognition and Understanding, 14-17, Dec. 1997, Pages 396 - 403

[4] R. Sarikaya and J.H.L. Hansen, *Improved Jacobian Adaptation for Fast Acoustic Model Adaptation in Noisy Speech Recognition,* ICSLP-2000: International Conf. Spoken Language Processing, vol. 3, pp. 702-705, Beijing, China, Oct. 2000

[5] ETSI draft standard doc. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 202 108 V1.1.2 (2000-04), http://pda.etsi.org/pda/queryform.asp, April 2000

[6] S. Young. HTK: *Hidden Markov Model Toolkit V1.5.* Cambridge Univ. Eng. Dept. Speech Group and Entropic Research Lab. Inc., Washington DC, Dec. 1993.