ROBUST SPEECH RECOGNITION FROM NOISE-TYPE BASED FEATURE COMPENSATION AND MODEL INTERPOLATION IN A MULTIPLE MODEL FRAMEWORK

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

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

ABSTRACT

Compared to multi-condition training (MTR), condition-dependent training generates multiple acoustic hidden Markov model sets each identified by a noisy environment and is known to perform substantially better for known noise types (included in training) while worse for unknown (untrained) noise types. This paper attempts to bridge the performance gap between known and unknown noise types by introducing a Minimum Mean-Square Error (MMSE) noise-type based compensation algorithm. On the basis of a modified Vector Taylor Series and the measurement of feature reliability as well as noise similarity, the MMSE estimation adapts the test features corrupted by the unknown noise type to the corresponding features corrupted by the known noise type. This method significantly improves the recognition performance for unknown noise types while maintaining the good performance for known noise types. Furthermore, in order to benefit directly from MTR, a model interpolation strategy is investigated which combines the MTR and the condition-dependent model sets. Both good performance and low computational cost are achieved by only interpolating the mixtures of each condition-dependent model state with the least weighted mixture in the corresponding MTR model state. The overall system gives promising results.

1. INTRODUCTION

It is well known that noisy environments significantly degrade the performance of an Automatic Speech Recognition (ASR) system, in particular when the system is trained on clean speech. The relatively low robustness against environmental noise has become one of the major obstacles for the widespread deployment of ASR technology.

To tackle this problem, multi-condition training (MTR) introduced in [1] was applied in training the acoustic Hidden Markov Model (HMM) set over a speech corpus corrupted by a number of noise types and Signal-to-Noise Ratios (SNR) likely to be encountered during use. The MTR method in general improves the ASR performances for both known (included in training) and unknown types of noise as compared to the clean training.

The performances for known noise types are further improved by introducing condition-dependent training strategies. In these strategies, multiple HMM sets are trained - each for a noise type as in [2] or for a combination of a noise type and a specific SNR value as in [3]. During recognition, model selection approaches are used to choose only one model set for recognition so that the extra computational cost introduced is as low as possible. Among these methods, it has been verified that the SNR and Noise Classification based Multiple Model Framework (SNC-MMF) performs best for the known types of noise [3].

However, it is worthwhile noting that training HMM models for all conditions (noise types) is normally impractical and condition-dependent training strategies must manage the problem with unknown noise types for which they often perform much worse than the MTR. Unfortunately, it is rare to see any research in the literature targeting this problem.

In this paper, the performance gap between known and unknown noise types is reduced by adapting the test feature corrupted by the unknown noise type to the feature corrupted by the known noise type corresponding to the selected SNC-MMF model set. This results in a noise-type based feature compensation method. For each Mel component, the method is implemented in two steps. In the first, the reliability and the noise similarity are measured to indicate the probability of the component being speech-dominant and the probability of the contained noise belonging to the known noise type, respectively. In the second step, the adaptation is performed by the Minimum Mean-Square Error (MMSE) estimation based on the measured reliability and noise similarity. Specifically, the Mel component which contains the same noise as the known noise type is unchanged to preserve information whereas the component containing the unknown noise type is either replaced directly by the known noise or adapted to the known-noise-corrupted noisy speech by a modified Vector Taylor Series (VTS) [4] depending on the reliability measurement. This method shows significant ASR performance improvement for the unknown noise types and maintains a good performance for the known types of noise.

Additionally, wanting to achieve further robustness to unknown noise types, this paper investigates model interpolation (MI) between the SNC-MMF and the MTR model states as the MTR models generally show good performance for unknown noise types. In [3], MI is implemented by linear combination of two model sets resulting in doubling the number of mixtures in each state and a doubling of the computational load. Instead of using all the mixtures in the MTR model set, this paper proposes to perform the MI between each SNC-MMF model state and the least weighted mixture within the corresponding MTR model state only. Compared against the old strategy, the models resulting from the new strategy achieve a similar recognition performance with a much lower computational complexity.

2. THE SNC-MMF FRAMEWORK

The SNC-MMF divides the noise corrupted training database

based on the type of additive noise and SNR value. A number of HMM model sets are then built - each for a combination of SNR value and noise type. This condition-dependent training leads to sharper Probability Density Functions (PDF) for each model set and thus ensures better discrimination for speech than the MTR. The efficiency of the ASR decoding is maintained by selecting only one model set according to the estimation of noise type and SNR value respectively from the noise classifier and the SNR estimator.

In [3], it has been experimentally verified that with only three model sets for each known noise type, significant improvement can be obtained for the known noise types as compared to the MTR method while the performance for the unknown noise types is lower, due to the training-test noise type mismatch. Thus, the challenge here is to improve the performance for unknown noise types while maintaining the good performance achieved for known noise types. Our solution is to mitigate the noise-type difference between the training and test through noise-type based feature compensation techniques, and this is achieved by performing the MMSE estimation in combination with the measurement of the reliability and noise similarity in each Mel component.

3. MMSE NOISE-TYPE BASED COMPENSATION

This section introduces the noise-type based compensation method in detail with the aim of adapting the test speech features corrupted by the unknown noise type (denoted as the source noise) into those corrupted by the known noise type (denoted as the target noise).

3.1 Measurement of the reliability and noise similarity

For each Mel component, the reliability *r* and noise similarity *nl* are first measured to identify whether the source speech signal is speech dominant (reliable) and whether the source and target noise types are the same.

3.1.1 Noise estimation for the source environment

Evaluating *r* and *nl* requires noise estimation for the source environment. Instead of using voice activity detection (VAD) based noise estimation as in [5] [6], the minimum statistical noise estimation (MSNE) [7] is adopted in this paper to acquire the noise estimation $\hat{N}_k(i)$ (the head denotes the estimated value) instantaneously for each logarithmical Mel component (*i*th) in each frame (*k*th) of the source noisy speech. With the observation that MSNE is still not accurate enough and only approximately tracks the averaged frame-by-frame noise changes, the noise is then modelled by a Gaussian distribution $\mathrm{N}(N_k(i); \hat{N}_k(i), \sigma^2_{N_Y}(i))$ where the mean is the noise estimation from MSNE and the variance $\sigma^2_{N_Y}(i)$ reflecting noise estimation errors is assumed stationary and is estimated from the first several non-speech frames of each utterance only.

3.1.2 Reliability

In this paper, the reliability is measured by the soft "missing data" mask [6]. Unlike the fuzzy mask used in [6] which adopts a sigmoid function, a more meaningful mask is obtained in terms of the above noise model by directly calculating the probability that noise energy is lower than the clean speech:

$$\begin{aligned} P^i{}_k(r) &= P(N_k(i) < X_k(i) \mid Y_k(i)) = P(2 * e^{N_k(i)} < e^{N_k(i)} + e^{X_k(i)} \mid Y_k(i)) \\ &= P(2 * e^{N_k(i)} < e^{Y_k(i)} \mid Y_k(i)) = P(N_k(i) < Y_k(i) - \ln(2) \mid Y_k(i)) \\ &= CDF(Y_k(i) - \ln 2; \hat{N}_k(i), \sigma^2_{N_Y}(i)) \end{aligned} \quad , \quad (1)$$

where $N_k(i)$, $X_k(i)$ and $Y_k(i)$ denote the *i*th logarithmical Mel components of the *k*th frame for source noise, clean speech and source noisy speech, respectively. The *CDF(x;a,b)* calculates the Gaussian cumulative distribution function with mean *a* and variance *b* at value *x* and can be efficiently implemented by a look-up table.

Eq.(1) assumes the noise and clean speech signals are additive in the Mel-spectral domain. Clearly, compared to the fuzzy mask, the probability based mask in Eq.(1) is more meaningful and free of sigmoid parameter tuning. In our implementation, the mask produced by Eq.(1) is further smoothed linearly within each frame to mitigate the potential errors in the produced mask.

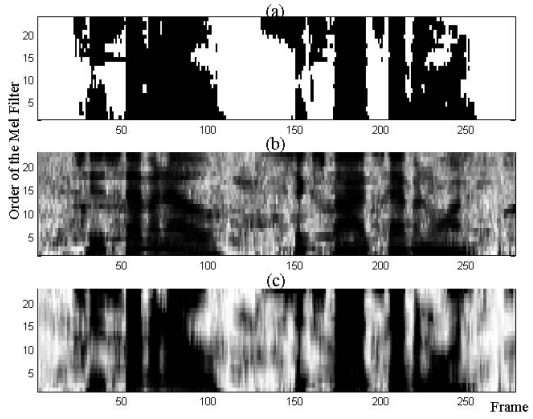

Fig.1 Comparisons among different soft masks: (a) *a priori* binary mask with a priori clean speech and noise; (b) the fuzzy mask as in [6] with the empirically determined optimal sigmoid parameters; (c) probability based soft mask

Fig.1 gives an example of the masks produced by *a priori* binary mask, by the fuzzy mask and by the probability based mask over the 10dB "restaurant" noisy speech sentence. Both the fuzzy mask and the probability based mask adopt the noise estimation from MSNE. It is observed that the probability based mask performs better than the fuzzy one especially in the noise dominated parts.

3.1.3 Noise similarity

Assuming the noise logarithmical Mel component Gaussian distributed, the noise similarity *nl* is evaluated by the Gaussian CDF function as:

$$\begin{aligned} P^i{}_k(nl) &= P(|N_Z(i) - \mu_{N_Z}(i)| > |\hat{N}_k(i) - \mu_{N_Z}(i)|) \\ &= 1 - |1 - 2 \cdot CDF(\hat{N}_k(i); \mu_{N_Z}(i), \sigma^2{}_{N_Z}(i))| \end{aligned} \quad , \quad (2)$$

where $N_Z(i)$ denotes the *i*th logarithmical Mel component of the target noise, and its mean $\mu_{N_Z}(i)$ and variance $\sigma^2{}_{N_Z}(i)$ can be estimated during training.

3.2 Estimation of the target Mel component

3.2.1 MMSE estimation

Given the source logarithmical Mel component $Y_k(i)$ which may contain unknown type of noise, MMSE is utilised to estimate the corresponding target component $Z_k(i)$ as:

$$\hat{Z_k}(i) = E[Z_k(i) \mid Y_k(i)] = \int Z_k(i) f(Z_k(i)|Y_k(i)) dZ_k(i) \,. \quad (3)$$

The conditional PDF $f(Z_k(i) \mid Y_k(i))$ can be extended based on the measured reliability and noise similarity:

$$\begin{aligned} f(Z_k(i) \mid Y_k(i)) &= P_k^i(nl)\delta(Z_k(i) - Y_k(i)) \\ &+ [1 - P_k^i(nl)]P_k^i(r)\mathrm{N}(Z_k(i); \mu^k{}_{Z|Y}(i), \sigma^2{}_{k,Z|Y}(i)) \\ &+ [1 - P_k^i(nl)][1 - P_k^i(r)]\mathrm{N}(Z_k(i); \mu_{N_Z}(i), \sigma^2{}_{N_Z}(i)) \end{aligned} \,, \quad (4)$$

where $\mu^k{}_{Z|Y}(i)$ and $\sigma^2{}_{k,Z|Y}(i)$ are respectively the mean and variance of the target speech conditioned on the observed source speech. Putting Eq.(4) into Eq.(3), the final estimation for the target noisy speech is given by:

$$\begin{aligned} \hat{Z_k}(i) &= P_k^i(nl)Y_k(i) + P_k^i(r)[1 - P_k^i(nl)]\mu_{Z|Y}(i) \\ &+ [1 - P_k^i(r)][1 - P_k^i(nl)]\mu_{N_Z}(i) \end{aligned} \,. \quad (5)$$

For different *nl-r* cases, Eq.(4) and (5) indicates different processing strategies: when the contained source noise is similar to the target noise type (the first term), the source and target noisy Mel components are treated as the same for the sake of maintaining the potential information in it; when noise type mismatch occurs, the conditional PDF of the noise-dominant component (the third term) is taken as the a priori distribution of the target noise type, and the PDF of the speech-dominate component (the second term) is assumed Gaussian and its mean value - $\mu_{Z|Y}$ - can be estimated later by a modified VTS.

3.2.2 Estimation of $\mu_{Z|Y}$ by VTS

To complete the calculation in Eq.(5), $\mu_{Z|Y}$ in this paper is estimated by the VTS algorithm. In the clean-training and noisy-test scenario, the original VTS [4] models the PDF of the clean speech logarithmical Mel-spectrum by a Gaussian Mixture Model (GMM). Given the test utterance and its noise estimation, it then uses MMSE to obtain the clean features for recognition. This is feasible for one-model cases where only one GMM is needed. However, since several noisy HMM model sets are employed in the SNC-MMF, it is not realistic to produce and store a number of GMM's each for a training acoustic environment. Instead, only the GMM for clean speech with 128 mixtures is generated and the nonlinear relationship between the source and target noisy speech in the logarithmical Mel domain are linearly approximated within each Gaussian mixture. The first-order VTS calculation is then modified as follows:

$$\begin{aligned} \mu^k{}_{Z|Y}(i) &= E[Z_k(i) \mid Y_k(i)] = \sum_m P(m|y) E[Z_k(i)|Y_k(i), m] \\ &= \sum_m P(m|y)[Y_k(i) - g(\mu_m(i), \overline{N(i)}) + g(\mu_m(i), \mu_{N_Z}(i))] \end{aligned} \quad (6a)$$

$$g(x, n) = \ln(1 + e^{n-x}) \quad (6b)$$

where μ_m is the mean value of the mth mixture in the clean speech GMM, and $g(x, n)$ the distortion function with the clean speech x and noise n. The conditional probability $P(m|y)$ is obtained the same as [4] by adapting the clean speech GMM to the noisy speech GMM for the test environment.

Instead of using the instant noise estimation from the MSNE, the noise estimation $\overline{N(i)}$ for the source speech signal is acquired from the first several non-speech frames so that the GMM adaptation and the calculation of the g functions only need to be fulfilled once for each utterance.

4. MODEL INTERPOLATION

It has been observed that the MTR models are generally more robust to unknown types of noise than the SNC-MMF models. In [3], a Full mixture MI (F-MI) method is introduced as follows:

$$f_I(O) = \alpha f_N(O) + (1 - \alpha) f_{MTR}(O) \,. \quad (7)$$

Given the observation O in Eq.(7), $f_I(O)$, $f_N(O)$ and $f_{MTR}(O)$ are the PDF's of an HMM state in the finally interpolated model set, the selected SNC-MMF noisy model set and the corresponding MTR model set, respectively. The interpolation factor α is empirically chosen as 0.4. This approach has proved to be effective for unknown noise types but inevitably doubles the number of Gaussian mixtures.

It is observed that Gaussian mixtures within each HMM state have different weights indicating different a priori probabilities. The larger the weight is, the more the mixture depends on the information contained in training data and is more sensitive to the mismatches between the test and training corpora. Thus, it is reasonable to expect the least weighted mixture in each MTR state contributes to the performance gain for unknown noise types most. The new interpolation strategy - Least weighted mixture MI (L-MI) - is adopted here by only interpolating the least weighted Gaussian mixture (PDF $f_{MTRL}(O)$) of each state in MTR model set with mixtures in the corresponding SNC-MMF state, i.e.

$$f_I(O) = \alpha f_N(O) + (1 - \alpha) f_{MTRL}(O) \,. \quad (8)$$

Obviously, this strategy introduces less number of mixtures in the interpolated model and can largely improve the computational efficiency as compared to F-MI.

5. EXPERIMENTS

The evaluations are conducted using the Aurora 2 database [8] which consists of connected English digits artificially corrupted by a number of additive noise types with SNR ranging from 20 to 0dB. The four noise types represented in test Set A ("Subway", "Babble", "Car" and "Exhibition") are treated as known noise types in the experiments while another four represented in Set B are as unknown.

Along with a 39-dimensional MFCC, the configuration of SNC-MMF is the same as [3]: three SNR-dependent HMM model sets are trained for each known noise type using SNR values close to 5dB, 10dB and 20dB, respectively; same to the MTR, each digit is modelled by 16 HMM states each with three Gaussian mixtures whereas the silence is modelled by 3 states each with 6 mixtures; for SNC-MMF model selection, the SNR estimation is conducted by a simple VAD-based SNR estimator, and the noise type is determined utterance by utterance based on the first 10 non-speech frames and a cepstral GMM based noise classifier.

Table 1 compares the averaged Word Error Rate (WER) performance over the two test sets for the SNC-MMF, the MTR and the proposed methods. The VTS, which uses $\mu^k_{Z|Y}(i)$ in Eq.(6) directly as the estimation of $Z_k(i)$, shows significant improvement for the unknown noise types but degrades the performance of the known noise types. The MMSE noise-type based compensation not only further improves the performance for the unknown noise types but also reclaims the VTS performance degradation for the known noise types. This demonstrates the necessity and effectiveness of using the reliability and noise similarity to protect the Mel components corrupted with known types of noise.

Table 1 Averaged WER (%) for different test sets

Methods \ Sets	Set A (Known)	Set B (Unknown)	Average
MTR	12.18	13.73	**12.96**
SNC-MMF	8.64	16.38	**12.51**
VTS	8.99	13.86	**11.43**
MMSE +VTS	8.63	12.69	**10.66**
F-MI	8.11	13.43	**10.77**
L-MI	8.32	13.61	**10.97**
F-MI+MMSE+VTS	8.51	11.45	**9.98**
L-MI+MMSE+VTS	8.53	11.79	**10.16**

Furthermore, a number of MI strategies are tested. As shown in Table 1, both F-MI and L-MI achieve high performance for both known and unknown noise types. Compared with F-MI, L-MI only degrades the recognition performance slightly but at the same time reduces the computational requirements a factor about 1/3 with the HMM model mixture number mentioned above. Finally, MMSE noise-type based compensation is combined with L-MI and further improves the overall performance which is very close to the results obtained by its combination with F-MI.

Fig.2 provides the detailed recognition performance for a number of unknown noise types in Set B. It demonstrates that our observations above are generally true for all the unknown noise types. In particular, the MMSE noise-type based compensation can in a large degree "assist" the VTS method to outperform the MTR for all the unknown noise types - especially for the "Restaurant".

6. CONCLUSIONS

The condition-dependent training shows a worse ASR performance than the MTR when unknown noise types occur during the recognition. This paper deals with this problem based on a recently proposed condition-dependent training framework - SNC-MMF. A feature-domain noise-type based compensation method is first presented to adapt the test speech features corrupted by the unknown noise type to the corresponding features corrupted by the known noise type. This method adopts the VTS and the measurement of the reliability and noise similarity to perform MMSE estimation and shows significant ASR performance improvement on the unknown noise types while maintaining the good performance for known noise types. Additionally, a new interpolation strategy between the SNC-MMF and MTR model sets is proposed and largely increases the efficiency of the full-mixture MI method by only interpolating each SNC-MMF model state with the least weighted mixture in the corresponding MTR HMM model state. The combination of the proposed methods is finally made and experiments over Aurora 2 indicate their promising performance for all the tested noise types.

7. 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.

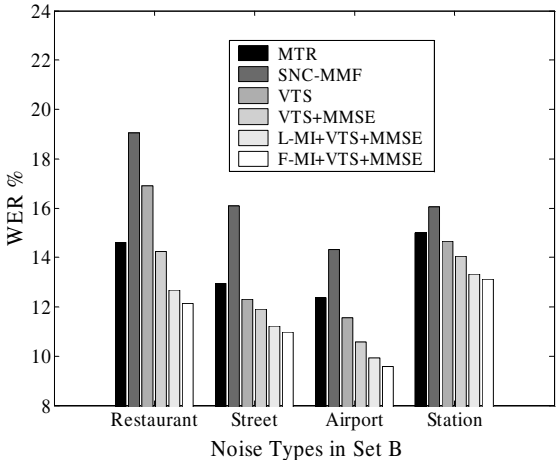

Fig.2 WER comparisons (averaged over different SNR's) for different unknown noise types in Set B

8. 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"* Proc. ICASSP '03, vol. 2, pp.113 -116, April 6-10, 2003

[3] H.Xu, Z.-H.Tan, P.Dalsgaard and B.Lindberg. "*Robust Speech Recognition Based on Noise and SNR Classification - a Multiple-Model Framework*". Proc. of INTERSPEECH 2005, Lisbon, Portugal, September, 2005

[4] P.J.Moreno. "*Speech Recognition in Noisy Environments* Ph.D. Thesis, ECE Department, CMU, May 1996.

[5] M.Cooke, P.Green, L.Josifovski and A.Vizinho, "*Robust automatic speech recognition with missing and unreliable acoustic data",* Speech Communication, Volume 34, Issue 3, June 2001

[6] J.Barker, L.Josifovski, M.Cooke and P.Green, "*Soft decisions in missing data techniques for robust automatic speech recognition*", Proceedings of ICSLP 2000, Beijing, 2000

[7] R. Martin, "*Noise power spectral density estimation based on optimal smoothing and minimum statistics*", IEEE Transactions on Speech and Audio Processing, vol.9, no.5, July 2001, pp.504 - 512

[8] H.G.Hirsch and D.Pearce, "*The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions*", ISCA ITRW ASR2000, Paris, France, September, 2000