Towards High-Accuracy Low-Cost Noisy Robust Speech Recognition Exploiting Structured Model

Jinyu Li JINYLI@MICROSOFT.COM
Li Deng DENG@MICROSOFT.COM
Dong Yu DONGYU@MICROSOFT.COM
Yifan Gong YGONG@MICROSOFT.COM
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

Abstract

It is well known that the distorted speech can be considered generated from the clean speech with the additive noise and the convolutive channel as In this paper, we present our recent study on using this structured model of physical distortion for robust automatic speech recognition. Three methods are introduced for joint compensation of additive and convolutive distortions (JAC), with different online computation costs. They are JAC model adaptation, GMM-based JAC model adaptation, and JAC feature enhancement. All these algorithms consist of two main steps. First, the noise and channel parameters are estimated using a nonlinear environment distortion model in the cepstral domain, and the vector-Taylor-series (VTS) linearization technique collectively. Second, the estimated noise and channel parameters are used to adapt the hidden Markov model (HMM) parameters or clean the distorted speech feature.

In the experimental evaluation using the standard Aurora 2 task, the proposed JAC algorithms all achieve around 89% accuracy using the clean-trained complex HMM backend, compare favorably over previously developed techniques. In the meanwhile, the JAC feature enhancement method has much smaller computation cost than the other two methods, and can be used as a high-accuracy low-cost noise robust front end. Detailed analysis on the experimental results shows that online updating all the noise and channel distortion parameters is critical to the success of our algorithms.

1. Introduction

Environment robustness in automatic speech recognition (ASR) remains an outstanding and difficult problem despite many years of research and investment [1]. The difficulty arises due to many possible types of distortions, including additive and convolutive distortions, which are not easy to predict accurately when developing the recognizers. As a result, the speech recognizer trained using clean speech often degrades its performance significantly when used under noisy environments if no compensation is applied. Different methodologies have been proposed in the past for environment robustness in speech recognition over the past two decades.

There are two main classes of approaches. In the first class, the distorted speech features are enhanced with advanced signal processing methods. The cleaned or enhanced speech features are then fed into the ASR system without dynamically changing the underlying acoustic models. Examples include the ETSI advanced front end (AFE) [2] and stereo-based piecewise linear compensation for environments (SPLICE) [3].

The other class of techniques operates on the model domain to adapt or adjust the model parameters so that the system becomes better matched to the distorted environment; Examples include maximum likelihood linear regression (MLLR) [4], parallel model combination (PMC) [5] and joint compensation of additive and convolutive distortions (JAC) [6][7]. The model-based techniques have shown better performance than the feature-based approaches [7][8].

One of the most powerful model adaption technologies is JAC which directly addresses the speech distortion exploiting a physical model. The distorted speech can be considered generated from the clean speech with the additive noise and the convolutive channel as As shown in [9], JAC has significant advantages over other model adaptation technologies such as MLLR, which adapts the acoustic model to the testing environment without imposing any constraints from the underlying physical distortion model.

Recently, we have developed several JAC-based model adaptation technologies such as online estimation of static and dynamic distortion parameters with vector Taylor series (VTS) expansion [9][10], phase-sensitive distortion model [11], and unscented transform for JAC [12]. Systems using these JAC methods have been demonstrated robust to distortions, and obtain very high recognition accuracy in noisy environments.

Although achieving high accuracy, JAC model adaptation methods share the same computational cost as other model adaptation technologies since it needs to adapt all the model parameters for every input utterance. This time-consuming requirement hinders JAC from being widely used, especially in large vocabulary continuous speech recognition (LVCSR) where the number of model parameters is large.

In this study, we investigate ways to reduce the online computation cost of JAC methods. Section 2 presents the physical distortion model and formularizes JAC model adaptation with VTS approximation. In Section 3, we provide the first solution which uses a clean-trained Gaussian mixture model (GMM) for online estimate of all distortion parameters, including static and dynamic parameters of noise and channel distortions. Section 4 presents feature-based JAC method which directly reduces distortions in the feature domain instead of adapting model parameters. In contrast to previous works [6], our proposed method utilizes our recently-developed technologies in JAC model adaptation by online estimating all the distortion parameters to ensure the quality of noise reduction. Our method can achieve almost the same accuracy as the model-based JAC method, with significantly reduced computational cost. The experiments are described in Section 5. We summarize our study and draw conclusions in Section 6.

2. JAC Model Adaptation Algorithm

Figure 1 shows a model for degraded speech with both noise (additive) and channel (convolutive) distortions. The observed distorted speech signal is generated from clean speech with noise and channel's impulse response according to

With discrete Fourier transformation (DFT), the equivalent relationship

can be established in the frequency domain, where k is the frequency-bin index in DFT given a fixed-length time window.

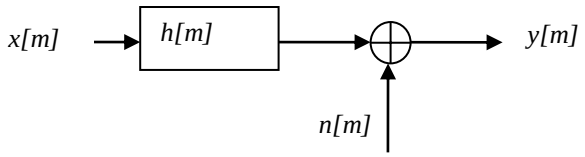

Figure 1. A model for acoustic environment distortion

The power spectrum of the distorted speech can then be obtained as

(1)

where denotes the (random) angle between the two complex variables and ( ).

It is noted that Eq. (1) is a general formulation for JAC. If is set to zero, Eq. (1) becomes

(2)

which is the formulation often used when power spectra [6] are adopted as the acoustic feature.

By taking logarithm and multiplying the non-square discrete cosine transform (DCT) matrix C to both sides of Eq. (2) for all the L Mel filter-banks, we obtain the nonlinear distortion model of

(3)

where , , , and are the clean speech, noise, channel, and distorted speech, respectively, in the cepstral domain. By taking the expectation on both sides of Eq. (3) and use vector Taylor series expansion (VTS), the static mean value of the distorted speech signal is

(4)

where

(5)

By noting,

(6)

(7)

we can derive the JAC-VTS adaption formulations for the k-th Gaussian in the j-th state as (following [9]):

(8)

(9)

(10)

(11)

(12)

(13)

We have proposed online estimation formulas for , , , , and in [10] which will not be repeated here.

The implementation steps of the JAC HMM adaptation algorithm described so far in this section and used in our experiments are plotted in Figure 2 and summarized in the following:

1. Read in a distorted speech utterance;
2. Set the channel mean vector to all zeros;
3. Initialize the noise mean vector and diagonal covariance matrix using the first and last *N* frames (speech-free) from the utterance using sample estimates;
4. Compute the Gaussian-dependent *G*(.) with Eq.(6), and adapt the HMM parameters with Eqs. (8)–(13);
5. Decode the utterance with the adapted HMM parameters;
6. Re-estimate noise and channel distortions using the above-decoded transcription;
7. Adapt the HMM parameters again with Eqs. (8)–(13);
8. Use the final adapted HMM model to decode the distorted speech feature and get output transcription.

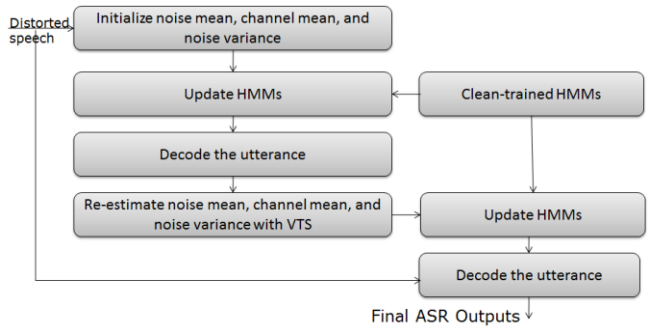

Figure 2. Flowchart of JAC model adaptation

3. GMM-Based JAC Model Adaptation Algorithm

The process in Section 2 is time consuming since it needs to adapt the HMM twice (step 4 and 7) and decode the utterance twice (step 5 and 8). If the HMM has large amount of parameters as in the LVCSR system, it makes JAC model adaptation unsuitable for real-time online adaptation despite the high accuracy it can achieve. To reduce the runtime cost, we can use clean-trained GMM for online estimation of all distortion parameters. The implementation steps for the GMM-based model adaptation algorithm are plotted in Figure 3 and described in the following:

1. Read in a distorted speech utterance;
2. Set the channel mean vector to all zeros;
3. Initialize the noise mean vector and diagonal covariance matrix using the first and last *N* frames (speech-free) from the utterance using sample estimates;
4. Compute the Gaussian-dependent *G*(.) with Eq.(6), and adapt the GMM parameters with Eqs. (8)–(13);
5. Re-estimate noise and channel distortions;
6. Adapt the HMM parameters with Eqs. (8)–(13);
7. Use the final adapted HMM model to decode the distorted speech feature and get output transcription.

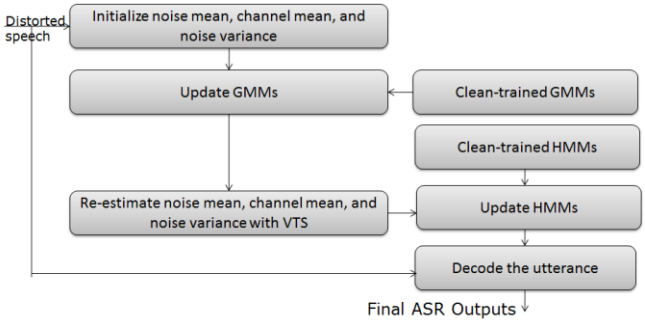

Figure 3: Flowchart of GMM-based JAC model adaptation

Usually, the GMM has much smaller number of parameters than the HMM. Therefore, the GMM adaption in step 4 has much less computation costs than the HMM adaptation. As a result, the GMM-based JAC adaptation method significantly reduces the runtime cost; with only one round full HMM model adaptation (step 6) and one round decoding (step 7).

4. JAC Feature Enhancement Algorithm

Although GMM-based JAC model adaptation can reduce the runtime cost from the JAC model adaptation in Section 2, the HMM adaptation in its step 6 is still time consuming in LVCSR system. In this section, we clean the distorted speech feature using JAC technology. The flowchart is in Figure 4.

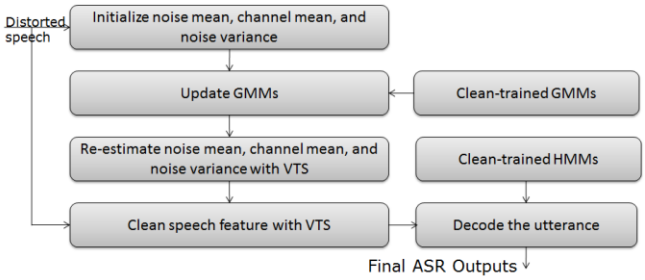

Figure 4: Flowchart of JAC feature enhancement

The following are the detailed implementation steps.

1. Read in a distorted speech utterance;

2. Set the channel mean vector to all zeros;
3. Initialize the noise mean vector and diagonal covariance matrix using the first and last N frames (speech-free) from the utterance using sample estimates;
4. Compute the Gaussian-dependent $G(.)$ with Eq.(6), and adapt the GMM parameters with Eqs. (8)–(13);
5. Online estimate noise and channel distortions;
6. Adapt the GMM parameters with Eqs. (8)–(13);
7. Use the final adapted GMM model to clean the distorted speech feature with Eq. (17) or Eq. (21);
8. Use the clean-trained HMM model to decoded the cleaned speech feature obtained in step 7 and get output transcription.

There is no more HMM adaptation step in this JAC feature enhancement technology. Given that the number of model parameters in GMM is much smaller than that in HMM, JAC feature enhancement has very reasonable runtime cost, and can be applied to LVCSR tasks. In the following, we discuss two JAC feature enhancement algorithms for step 7.In general, we can use the minimum mean square error (MMSE) method to get the estimate of clean speech

(14)

Suppose the clean-trained GMM is denoted as

together with Eq. (3), we have

(15)

Here,

(16)

with updated in step 6 using Eqs. (8)–(13).

If we use as the 0th-order approximation of , then

(17)

This formulation was first proposed in [6], and we denote it as JAC-0.In [13], another solution was proposed when expanding Eq. (3) with 1st-order VTS:

With some derivations, we can have

(18)

with

(19)

Assume the noise and clean speech are independent, we get

(20)

The final solution denoted as JAC-1 is

(21)

Note that although Eqs. (17) and (21) were also proposed in [6] and [13], they are not widely used due to the inferior accuracies than the model-based JAC methods. The key of JAC-based feature enhancement is to get a reliable estimation of noise and channel distortion parameters, and accurately calculate the Gaussian occupancy probability in Eq. (16). In [6] and [13], only static noise and channel mean vectors are estimated. In contrast, we propose to online update all the distortion parameters [10]. After updating both the static and dynamic model parameters with the online distortion estimations, we can have a more accurate estimation of the Gaussian occupancy probability. As shown in the following experiment section, this is critical to the success of JAC-based feature enhancement.

5. Experiments and Discussions

The effectiveness of different JAC algorithms presented in Section 2, 3, and 4 has been evaluated on the standard Aurora 2 task [14] of recognizing digit strings in noise and channel distorted environments. The clean training set, which consists of 8440 clean utterances, is used to train the baseline maximum likelihood estimation (MLE) HMMs. The test material consists of three sets of distorted utterances. The data in set-A and set-B contain eight different types of additive noise, while set-C contain two different types of noise plus additional channel distortion. Each type of noise is added into a subset of clean speech utterances, with seven different levels of signal to noise ratios (SNRs). This generates seven subgroups of test sets for a specified noise type, with clean, 20db, 15db, 10db, 5db, 0db, and -5db SNRs. The

baseline experiment setup follows the standard script provided by ETSI [14], including the complex "backend" of HMMs trained using the HTK toolkit.

In the complex backend provided by [14], there are 11 whole-digit HMMs, one for each of the 11 English digits, including the word "oh". Each HMM has 16 states, with simple left-to-right structure and no skips over states. Each state is modeled by a Gaussian mixture model (GMM) with 20 Gaussians. All HMM's covariance matrices are diagonal. In addition, there are one "sil" and one "sp" model. The "sil" model consists of 3 states, and each state is modeled by a GMM with 36 Gaussians. The "sp" model has only one state and is tied to the middle state of the "sil" model. The total number of Gaussians in the HMM is 3628. We also train a single GMM with 552 Gaussians for the algorithms in Section 3 and 4.

The features are 13-dimensional MFCCs, appended by their first- and second-order time derivatives.

The cepstral coefficient of order 0 is used instead of the log energy in the original script. This gives a baseline of 61.51% Accuracy (Acc).

The JAC algorithms presented in this paper are then used to adapt the above MLE HMMs or to clean the distorted feature utterance by utterance for the entire test set (Sets-A, B, and C). The detailed implementation steps described in Section 2, 3, and 4 are used in the experiments. We use the first and last N=20 frames from each utterance for initializing the noise means and variances.

Table 1 summarizes the recognition accuracy of the baseline, the JAC model adaptation algorithm described in Section 2, the GMM-based JAC model adaptation algorithm presented in Section 3, and the JAC feature enhancement algorithm discussed in Section 4. JAC model adaptation obtained the best accuracy of 89.99%.

We evaluate GMM-based JAC model adaptation with two setups. The first is to combine all the 3628 Gaussians in the HMM model into a GMM. The other setup is to use the single GMM with 552 Gaussians. Both setups give very similar accuracy: a little higher than 89%. This means that 552 Gaussians are good enough for this task using GMM-based JAC model adaptation technique. The accuracy gap between the algorithms in Section 2 and 3 is due to the high-quality decoded transcription in Step 5 of JAC model adaptation. That transcription guides the online noise and channel estimations with limited useful Gaussians. In contrast, GMM-based JAC adaptation uses all the Gaussians for online estimation.

The accuracies of two JAC feature enhancement algorithms are also compared in Table 1. Both methods use the GMM with 552 Gaussians. JAC-0 gets 88.61% accuracy which is very close to the 89.06% accuracy achieved by GMM-based JAC model adaptation. In contrast, JAC-1 only got 86.08% accuracy, making JAC-0 a better choice of high-accuracy low-cost noise robust algorithm.

Table 1. Recognition accuracy of the baseline, JAC model adaptation, GMM-based JAC model adaptation, and JAC feature enhancement algorithms

Algorithm	Accuracy
Baseline	61.51%
JAC model adaptation	89.99%
GMM-based JAC model adaptation (Combine all the Gaussians in HMM to form GMM)	89.13%
GMM-based JAC model adaptation (GMM with 552 Gaussians)	89.06%
JAC-0 feature enhancement	88.61%
JAC-1 feature enhancement	86.08%

To examine the effect of individual contributions of JAC feature enhancement algorithms, we conducted several experiments incrementally and summarized the results in Table 2. If steps 5 and 6 in Section 4 are skipped, JAC-0 obtains 86.12% accuracy since only the initial noise and channel estimation was used to clean the distorted speech feature,. If we use all steps but only update noise and channel mean parameters as in [6], the accuracy is increased to 86.72% for JAC-0. After updating all the distortion parameters (mean and variance for both static and dynamic features), the accuracy of JAC-0 can be boosted to 88.61%. The similar observation can be made for the JAC-1 algorithm.

Table 2. Recognition accuracy of JAC feature enhancement algorithms with different options

	JAC-0 feature enhancement	JAC-1 feature enhancement
No online distortion update	86.12%	84.63%
Online update noise and channel mean parameters only	86.72%	84.69%
Online update all distortion parameters	88.61%	86.08%

6. Conclusions

In this paper, we have presented a way to get a high-accuracy low-cost noise robust ASR system. Starting from the JAC model adaptation which enjoys high accuracy at high computational cost, we improve the efficiency by using GMM-based JAC adaptation and JAC-feature enhancement. All these technologies are built on the basis of the speech distortion physical model. By online estimating all the noise and channel distortion

parameters, we can obtain high accuracies for all the presented technologies.

In the experimental evaluation using the standard Aurora 2 task, the JAC-0 feature enhancement algorithm achieved 88.61% accuracy using the clean-trained complex HMM backend. The JAC feature enhancement algorithm achieves almost the same accuracy as the GMM-based JAC model adaptation, with much less online computational cost. This enables us to deploy JAC feature enhancement method to the LVCSR tasks.

Several research issues will be addressed in the future. First, this paper studied 0th order and 1st order expansion for feature enhancement. In [15], high order expansion is used to make the approximation accurate. We will explore this direction in the future. Second, the success of JAC feature enhancement methods heavily relies on whether we can have a good adapted GMM. In current study, we only use the standard VTS technology to update GMMs while we have shown that phase-sensitive distortion VTS and unscented transform technologies can help to improve the modeling quality [11][12]. We will apply these technologies in the next step.

References

[1] Peinado, A. and Segura, J. *Speech Recognition over Digital Channels --- Robustness and Standards*. John Wiley and Sons Ltd (West Sussex, England), 2006.

[2] Macho, D., Mauuary, L., Noe, B., Cheng, Y. M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., and Saadoun, F. Evaluation of a noise-robust DSR front-end on Aurora databases. in *Proc. ICSLP*, pp. 17–20, 2002.

[3] Deng, L., Acero, A., Plumpe, M., and Huang, X. Large vocabulary speech recognition under adverse acoustic environments. in *Proc. ICSLP*, Vol.3, pp.806-809, 2000.

[4] Leggetter, C. J. and Woodland, P. C. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer, Speech, Lang,*, Vol. 9, No. 2, pp. 171–185, 1995.

[5] Gales, M. J. F. and Young, S. An improved approach to the hidden Markov model decomposition of speech and noise. in *Proc. ICASSP*, Vol. I, pp. 233–236, 1992.

[6] Moreno, P. *Speech Recognition in Noisy Environments*. PhD. Thesis, Carnegie Mellon University, 1996.

[7] Gong, Y. A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition. *IEEE Trans. Speech and Audio Proc.*, Vol. 13, No. 5, pp. 975-983, 2005.

[8] Acero, A., Deng, L., Kristjansson, T., and Zhang, J. HMM adaptation using vector Taylor series for noisy speech recognition. in *Proc. ICSLP*, Vol.3, pp. 869-872, 2000.

[9] Li, J., Deng, L., Yu, D., Gong, Y., and Acero. A. High-performance HMM adaptation with joint compensation of additive and convolutive distortions. in *Proc. IEEE ASRU*, 2007.

[10] Li, J., Deng, L., Yu, D., Gong, Y., and Acero. A. A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions *Computer Speech and Language*, vol. 23, pp. 389-405, 2009.

[11] Li, J., Deng, L., Yu, D., Gong, Y., and Acero, A. HMM adaptation using a phase-sensitive acoustic distortion model for environment-robust speech recognition. in *Proc. ICASSP*, April 2008.

[12] Li, J., Yu, D., Gong, Y., and Li Deng. Unscented transform with online distortion estimation for HMM adaptation. in Interspeech 2010, September 2010.

[13] Stouten, V., Hamme, H. V., Demuynck, K. and Wambacq, P. Robust speech recognition using model-based feature enhancement. In *Proc. European Conference on Speech Communication and Technology*, pp. 17-20, 2003.

[14] Hirsch, H. G. and Pearce, D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. in *Proc. ISCA ITRW ASR*, 2000.

[15] Du, J. and Huo, Q. A feature compensation approach using high-order vector Taylor series approximation of an explicit distortion model for noisy speech recognition. *IEEE Trans. Audio, Speech and Lang. Proc.*, 2011.