NOISE MODEL TRANSFER USING AFFINE TRANSFORMATION WITH APPLICATION TO LARGE VOCABULARY REVERBERANT SPEECH RECOGNITION

Takuya Yoshioka, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper considers using the feature enhancement approach for automatic recognition of speech corrupted by severely nonstationary noise, caused for example by interfering talkers and inter-frame distortion induced by reverberation. In particular, we focus on the issue of feature-domain noise model estimation and investigate a recently proposed approach, called noise model transfer (NMT), for estimating the rapidly changing noise model parameter values. Based on the fact that noise spectral changes can be detected more easily in the power spectrum domain than in the feature domain, NMT estimates the noise model parameter values for each time frame by using both observed feature vectors and noise power spectral estimates, on the assumption that a separate noise power spectrum estimator is available. This is achieved by finding the best transformation that maps the power spectra onto the noise model parameter space in the maximum likelihood sense. Whereas the transformation was previously modeled using a bias vector, this paper employs a more flexible affine transformation model. The results of 20,000-word reverberant speech recognition experiments show the advantage of the affine transformation model.

Index Terms— Robust speech recognition, nonstationary noise, reverberation, vector Taylor series, noise model transfer

1. INTRODUCTION

Noise robustness has been one of the main topics in the automatic speech recognition (ASR) research, and its importance appears to be growing with the rapid spread of ASR technology. Of the various noise robustness approaches, this paper focuses on feature enhancement methods, which attempt to estimate a sequence of clean feature vectors based on the corresponding corrupted feature vector sequence. One exemplary method called vector Taylor series (VTS) enhancement exploits models of clean and noise feature vectors [1,2] to achieve this estimation task. The parameters of the clean speech model are trained in advance while those of the noise model are estimated in the recognition phase based on the observed data.

One of the unsolved challenges as regards feature enhancement is to achieve accurate noise model estimation under 'ultra-nonstationary' noise conditions, which are typically created by interfering talkers and reverberation. Most conventional methods employ a noise model consisting of a single Gaussian probability distribution of additive noise feature vectors and a constant bias representing a convolutional noise feature vector. The model parameters are usually estimated on the assumption that they are invariant [1, 2] or change slowly with time [3–5]. However, this assumption hinders modeling and the cancellation of the ultra-nonstationary noise[1].

To solve this problem, we proposed an approach called *noise model transfer* (NMT) [6] by extending the idea described in [7]. Unlike conventional methods, this approach can exploit both observed feature vectors, representing spectral envelopes, and other signal properties, such as harmonic structures and phase spectra, that are usually discarded during feature extraction but that are useful for separating nonstationary noise from target speech. In NMT, we assume the availability of a reliable noise power spectrum estimator that can quickly track noise changes by leveraging such additional signal properties. Then, we calculate the optimal transformation that maps each of the estimated frame-specific noise power spectra to the space of the noise model parameters in the maximum likelihood sense. The transformation can be modeled in many different ways, and a bias transformation model was employed in [6]. Even with the simple bias transformation model, NMT removed a quantity of recognition errors in both meeting speech and reverberant speech recognition tasks.

This paper presents an algorithm for performing NMT using an affine transformation, which has more adjustable parameters and thus is more flexible than the bias transformation model. Specifically, we hypothesize an affine transformation that projects the feature-domain representation of each noise power spectrum estimate that is obtained using the separate noise power spectrum estimator onto the mean vector space of an additive noise model. The affine transformation matrix is optimized jointly with the additive noise covariance matrix and the convolutional noise feature vector to maximize the parameter likelihood. The resulting algorithm is applied to reverberant speech recognition by regarding late reverberation as nonstationary additive noise [8]. This paper also expands the experiment beyond that described in our previous report [6]: whereas we previously performed a reverberant digit recognition test, here we use a 20,000-word reverberant speech recognition task under both clean and multi-style training conditions.

2. REVIEW OF NOISE MODEL TRANSFER

This section reviews the concept of NMT after briefly looking at the VTS enhancement method, which is used along with NMT.

2.1. Vector Taylor series for feature enhancement

VTS is an approach to noise robust speech recognition. The goal is to correctly transcribe an observed speech signal that is corrupted by environmental noise. The observed signal, denoted by $y(s)$, is generated as

$$y(s) = h(s) \circledast x(s) + n(s), \tag{1}$$

[1]The dynamic noise adaptation proposed in [5] can deal with large sporadic changes in noise characteristics. However, it seems ineffective for describing continuous changes that are characteristic of interfering speech and reverberation noise.

where $x(s)$ is a clean speech signal, $n(s)$ an independent additive noise signal, $h(s)$ an impulse response describing convolutional noise, and $\circledast$ denotes a linear convolution operator. While (1) describes the relationship between the clean and corrupted speech signals in the time domain, the relationship is approximately represented by the following equation in the feature domain (assuming the use of logarithmic mel-frequency spectral representation):

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{h} + \log(1 + \exp(\boldsymbol{n}_t - \boldsymbol{x}_t - \boldsymbol{h})), \tag{2}$$

where $\boldsymbol{y}_t$, $\boldsymbol{x}_t$, $\boldsymbol{h}$, and $\boldsymbol{n}_t$ are the vectors of the logarithmic mel-frequency spectral features extracted from $y(s)$, $x(s)$, $h(s)$, and $n(s)$, respectively, at the tth short time frame[2]. The right hand side of (2) is called the mismatch function and is hereafter represented as $f(\boldsymbol{x}_t, \boldsymbol{n}_t, \boldsymbol{h})$.

Given the corrupted feature vector $\boldsymbol{y}_t$, the VTS method estimates the underlying clean feature vector $\boldsymbol{x}_t$ by leveraging models of clean and additive noise feature vectors so that a subsequent speech recognizer can receive feature vectors that are as clean as possible. The clean speech model is often represented in the form of a Gaussian mixture model (GMM) with diagonal covariance matrices as follows:

$$p_{\mathrm{x}}(\boldsymbol{x}_t) = \sum_{k=1}^{K} \pi_k f_{\mathrm{N}}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{\mathrm{x}}, \mathrm{diag}(\boldsymbol{\sigma}_k^{\mathrm{x}})), \tag{3}$$

where K is the number of Gaussians, $\mathrm{diag}(\boldsymbol{x})$ represents the diagonal matrix with the elements of vector $\boldsymbol{x}$ on its leading diagonal, and $f_{\mathrm{N}}(\cdot)$ is the probability density function (pdf) of the multivariate normal distribution. On the other hand, the additive noise model takes the form of a single Gaussian pdf given by

$$p_{\mathrm{n}}(\boldsymbol{n}_t; \theta_t) = f_{\mathrm{N}}(\boldsymbol{n}_t; \boldsymbol{\mu}_t^{\mathrm{n}}, \mathrm{diag}(\boldsymbol{\sigma}_t^{\mathrm{n}})), \tag{4}$$

where $\theta_t = \{\boldsymbol{\mu}_t^{\mathrm{n}}, \boldsymbol{\sigma}_t^{\mathrm{n}}\}$ with $\boldsymbol{\mu}_t^{\mathrm{n}}$ and $\boldsymbol{\sigma}_t^{\mathrm{n}}$ being the mean vector and the leading diagonal of the covariance matrix, respectively. Note that the additive noise model parameters depend on frame index t to account for nonstationary noise. The clean speech model parameters $\{\pi_k, \boldsymbol{\mu}_k^{\mathrm{x}}, \boldsymbol{\sigma}_k^{\mathrm{x}}\}_{1 \le k \le K}$ are learned from training data. On the other hand, the time series of the additive noise model parameters $(\theta_t)_{1 \le t \le T}$ has to be estimated jointly with the convolutional noise feature vector $\boldsymbol{h}$ based on the observed data $(\boldsymbol{y}_t)_t$ (we hereafter omit the range of index values when representing sets and sequences). With the above clean speech and additive noise models, a minimum mean square error (MMSE) estimate of the clean feature vector $\boldsymbol{x}_t$ can be calculated by linearizing the mismatch function $f(\cdot)$ around the mean vectors of the clean speech and additive noise models (see [1] for details).

2.2. Noise model transfer (NMT)

NMT is an approach for estimating the additive noise model parameters $(\theta_t)_t$ and the convolutional noise feature vector $\boldsymbol{h}$. This novel approach takes advantage of a separate noise power spectrum estimator that can promptly detect changes in noise characteristics. This is different from existing methods, which estimate the noise model parameters directly from the observed feature vectors $(\boldsymbol{y}_t)_t$. The underlying philosophy is that tracking very nonstationary noise is difficult to achieve in the feature domain and should be done in the time or power spectrum domain. This is because the feature vectors retain only the spectral envelope information while the other signal properties, which are usually discarded during feature extraction, provide useful information for noise estimation. Thus, an NMT-based noise model estimator uses both the additive noise power spectral sequence, $(\hat{\boldsymbol{N}}_t)_t$, produced by the noise power spectrum estimator and the corrupted feature vector sequence $(\boldsymbol{y}_t)_t$.

With the NMT approach, we hypothesize a transformation $z(\cdot)$ that maps each additive noise power spectrum estimate $\hat{\boldsymbol{N}}_t$ into the parameter value set of the additive noise model as follows:

$$\theta_t = z(\hat{\boldsymbol{N}}_t; \phi), \tag{5}$$

where parameter set ϕ characterizes the transformation. Then we jointly estimate the transformation parameters and the convolutional noise feature vector to maximize the likelihood as

$$(\hat{\phi}, \hat{\boldsymbol{h}}) = \underset{(\phi, \boldsymbol{h})}{\mathrm{argmax}} \prod_{t=1}^{T} p_{\mathrm{y}}(\boldsymbol{y}_t; z(\hat{\boldsymbol{N}}_t; \phi), \boldsymbol{h}), \tag{6}$$

instead of directly estimating the noise model parameters. Function $p_{\mathrm{y}}(\cdot)$ is the pdf of the observed feature vector, which is calculated by combining the clean speech and noise models. One desirable property of NMT is that we can allow for large changes in the additive noise feature vector model parameters while keeping the number of adjustable parameters small. Our previous study [6] employed the bias transformation model given by

$$\boldsymbol{\mu}_t^{\mathrm{n}} = d(\boldsymbol{N}_t) + \boldsymbol{b}; \quad \boldsymbol{\sigma}_t^{\mathrm{n}} = \boldsymbol{c}, \tag{7}$$

where $\boldsymbol{b}$ and $\boldsymbol{c}$ are the transformation parameters, i.e., $\phi = \{\boldsymbol{b}, \boldsymbol{c}\}$. Function $d(\cdot)$ denotes an operator that calculates a feature vector from a power spectrum. Specifically, $d(\boldsymbol{X}) = \log(\boldsymbol{W}\boldsymbol{X})$, where $\boldsymbol{W}$ is a mel-frequency filter-bank matrix.

3. NMT USING AFFINE TRANSFORMATION

This paper investigates the following affine transformation model:

$$\boldsymbol{\mu}_t^{\mathrm{n}} = \boldsymbol{A} d(\boldsymbol{N}_t) + \boldsymbol{b}; \quad \boldsymbol{\sigma}_t^{\mathrm{n}} = \boldsymbol{c}, \tag{8}$$

where $\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ comprise the transformation parameter set ϕ. This model is more flexible than the bias model and is expected to yield better noise model parameter values. The first equation of (8) is equivalent to the following equation:

$$\boldsymbol{\mu}_t^{\mathrm{n}} = \boldsymbol{B}\boldsymbol{d}_t, \tag{9}$$

where $\boldsymbol{B} = [\boldsymbol{A}, \boldsymbol{b}]$ and $\boldsymbol{d}_t = [d(\boldsymbol{N}_t)^T, 1]^T$. Hence, we jointly optimize the affine transformation matrix $\boldsymbol{B}$, the additive noise covariance vector $\boldsymbol{c}$, and the convolutional noise feature vector $\boldsymbol{h}$ to maximize the likelihood as shown in (6), where the transformation parameter ϕ in (6) is given by $\phi = \{\boldsymbol{B}, \boldsymbol{c}\}$.

We employ the expectation-maximization (EM) framework to achieve the maximum likelihood estimation task given by (6). We derive an optimization algorithm by regarding the Gaussian index of the clean speech model and the additive noise feature vector as the latent variables of the EM. Thus, at the E-step, we calculate the following two quantities for each possible combination of t and k values using the current parameter value set $\{\hat{\phi}, \hat{\boldsymbol{h}}\}$:

1. the posterior probability $\gamma_{t,k}$ of the underlying clean feature vector $\boldsymbol{x}_t$ being generated from the kth component of the clean speech GMM;

[2]When deriving (2) from (1), we assumed $h(s)$ to be sufficiently short so that it could be approximated as multiplicative noise in the short time Fourier transform (STFT) domain. As discussed in Section 4, we deal with reverberation, which is represented by long impulse responses, by regarding the late reverberation, constituting a large part of the interferences caused by reverberation, as nonstationary additive noise.

2. the posterior pdf $p_{\mathrm{n|y,k}}(\boldsymbol{n}_t|\boldsymbol{y}_t,k;\hat{\phi},\hat{\boldsymbol{h}})$ of the true additive noise feature vector $\boldsymbol{n}_t$. This pdf is expressed by a Gaussian distribution. We denote its mean vector and covariance matrix by $\boldsymbol{\mu}_{t,k}^{\mathrm{n|y}}$ and $\boldsymbol{\Sigma}_{t,k}^{\mathrm{n|y}}$, respectively.

See Appendix for the formulae for caluculating these posteriors. Then, at the M-step, we calculate a set of updated parameter values by maximizing the following auxiliary function:

$$Q(\phi,\boldsymbol{h}) = \sum_{t,k}\left\langle \log p_{\mathrm{y|n,k}}(\boldsymbol{y}_t|\boldsymbol{n}_t,k;\boldsymbol{h}\backslash\boldsymbol{\mu}_l^{\mathrm{x}},\hat{\phi},\hat{\boldsymbol{h}}) + \log p_{\mathrm{n}}(\boldsymbol{n}_t;\phi)\right\rangle, \quad (10)$$

where the operator $\langle\cdot\rangle$ takes the expectation over $\boldsymbol{n}_t$ with respect to the posterior distribution of $\boldsymbol{n}_t$ calculated at the E-step. Function $p_{\mathrm{y|n,k}}(\cdot)$ represents the conditional pdf of the corrupted feature vector given the additive noise feature vector and the component index of the clean speech GMM. As shown in Appendix, $p_{\mathrm{y|n,k}}(\cdot)$ is a Gaussian distribution, whose mean vector $\boldsymbol{\mu}_{t,k}^{\mathrm{y|n}}$ and covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}_{t,k}^{\mathrm{y|n}})$ are obtained by linearizing the mismatch function using the zeroth and first-order terms of its Taylor series. The backslash that appears in the notation $p_{\mathrm{y|n,k}}(\cdot)$ means that the Taylor series is calculated around the variables placed after the backslash. As the involved pdfs are all Gaussians, the parameter values maximizing the auxiliary function can be easily obtained as[3]

$$\hat{\boldsymbol{B}}^{(\mathrm{new})} = \boldsymbol{R}_{\mu\mathrm{d}}\boldsymbol{R}_{\mathrm{dd}}^{-1}; \quad \hat{\boldsymbol{c}}^{(\mathrm{new})} = \frac{1}{T}\left(\boldsymbol{R}_{\mu\mu} - \hat{\boldsymbol{B}}^{(\mathrm{new})}\boldsymbol{R}_{\mu\mathrm{d}}^{T}\right); \quad (11)$$

$$\hat{\boldsymbol{h}}^{(\mathrm{new})} = \frac{\sum_{t,k}\gamma_{t,k}\boldsymbol{h}_{t,k}^{\top}}{\sum_{t,k}\gamma_{t,k}\boldsymbol{h}_{t,k}^{\perp}} + \hat{\boldsymbol{h}}, \quad (12)$$

using the following definitions:

$$\boldsymbol{R}_{\mathrm{dd}} = \sum_{t,k}\gamma_{t,k}\boldsymbol{d}_t\boldsymbol{d}_t^{T}; \quad \boldsymbol{R}_{\mu\mathrm{d}} = \sum_{t,k}\gamma_{t,k}\boldsymbol{\mu}_{t,k}^{\mathrm{n|y}}\boldsymbol{d}_t^{T}; \quad (13)$$

$$\boldsymbol{R}_{\mu\mu} = \sum_{t,k}\gamma_{t,k}\left(\boldsymbol{\mu}_{t,k}^{\mathrm{n|y}}\left(\boldsymbol{\mu}_{t,k}^{\mathrm{n|y}}\right)^{T} + \boldsymbol{\Sigma}_{t,k}^{\mathrm{n|y}}\right); \quad (14)$$

$$\boldsymbol{h}_{t,k}^{\top} = \hat{\boldsymbol{g}}_{t,k}\left(\boldsymbol{y}_t - \hat{\boldsymbol{f}}_{t,k} - \left(1-\hat{\boldsymbol{g}}_{t,k}\right)\left(\boldsymbol{\mu}_{t,k}^{\mathrm{n|y}} - \hat{\boldsymbol{B}}\boldsymbol{d}_t\right)\right); \quad \boldsymbol{h}_{t,k}^{\perp} = \frac{\hat{\boldsymbol{g}}_{t,k}}{\boldsymbol{\sigma}_{t,k}^{\mathrm{y|n}}}, \quad (15)$$

where $\hat{\boldsymbol{f}}_{t,k} = f\left(\boldsymbol{\mu}_k^{\mathrm{x}}, \hat{\boldsymbol{B}}\boldsymbol{d}_t, \hat{\boldsymbol{h}}\right)$ and $\hat{\boldsymbol{g}}_{t,k} = g\left(\boldsymbol{\mu}_k^{\mathrm{x}}, \hat{\boldsymbol{B}}\boldsymbol{d}_t, \hat{\boldsymbol{h}}\right)$ with function $g(\cdot)$ being the partial derivative of the mismatch function $f(\cdot)$ with respect to the clean feature vector. The multiplication and division operations are performed on an element-by-element basis.

4. REVERBERANT SPEECH RECOGNITION USING NMT

Now, we consider applying NMT to single-microphone reverberant speech recognition. The reverberation seriously degrades the recognition performance. This problem has already been tackled by some researchers, yielding solutions ranging from signal dereverberation [10–12] to acoustic model adaptation [13–16]. A feature enhancement method for reverberant speech recognition is proposed in [17]. A comprehensive review of the reverberant speech recognition techniques is provided in [8].

To apply NMT to reverberant speech recognition, we regard late reverberation as nonstationary additive noise as discussed below. It is known that a room impulse response, which describes the change in a speech signal caused by reverberation, can be divided into three portions that are called direct sound, early reflections, and late reverberation [18]. The late reverberation consists of myriad reflections that arrive after the early reflections, which typically disappear within 50 ms. Since the autocorrelation of a clean speech signal tends to vanish at lags greater than 50 ms, the convolution of the late reverberation and the clean speech signal, which is also called the late reverberation, is almost uncorrelated with the direct sound. Furthermore, by definition, the early reflections are represented by a short impulse response. Therefore, we can reasonably use the conventional mismatch function given by (2) to describe the impact of reverberation in the feature domain, which allows us to exploit the widely used VTS method. This is one advantage of the reverberant feature enhancement method described here over the method proposed in [17], which employs complex models of speech and reverberation dedicated to reverberant feature enhancement. However, please note that we do not intend to contrast the proposed method and existing reverberant feature enhancement methods. The aim of this paper is to compare the bias transformation model and the affine transformation model. Although, we consider reverberant speech recognition as a test bed, NMT is applicable to a wide range of tasks including meeting speech recognition [6] and microphone-array speech recognition [7].

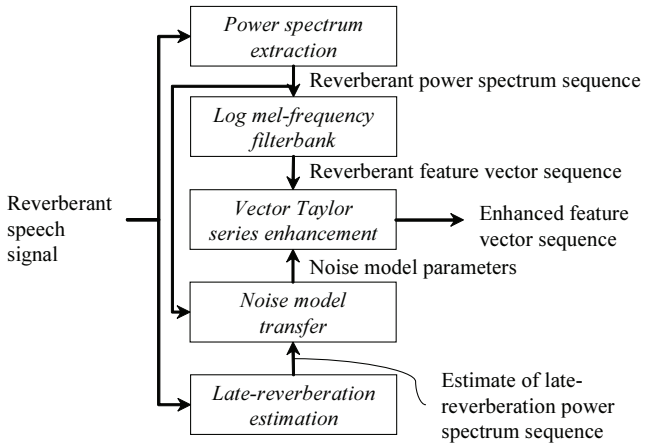

Fig. 1. Processing and data flow diagram of NMT-based reverberant feature enhancement.

Based on the development described above, we can construct a complete NMT-based reverberant feature enhancement algorithm. The algorithm is explained below using the processing and data flow diagram shown in Fig. 1.

1. *Power spectrum extraction:* The short-time Fourier transform of a reverberant speech signal captured by a microphone is calculated. Then, the squared magnitudes of the resulting Fourier coefficients are computed to obtain the short-time power spectra of the reverberant speech signal.
2. *Log mel-frequency filterbank:* The reverberant power spectra are fed into a mel-frequency filter bank, and the logarithms of the filter bank outputs are calculated to obtain the logarithmic mel-frequency spectral feature vectors of the reverberant speech signal.
3. *Late-reverberation estimation:* The late-reverberation power spectra are calculated by using the method proposed in [19]. This method estimates each short-time power spectrum of

[3] Note that the set of the presented update formulae does not necessarily increase the likelihood because the center of the Taylor series expansion, which depends on the parameter values, changes from iteration to iteration. A back-off method can be employed to ensure the likelihood increase [9].

Table 1. WER results of reverberant speech recognition experiments. The baseline clean speech WER was 10.2%.

	Clean training	Multistyle training
Reverberant	93.5%	44.8%
Bias NMT	55.9 %	35.1%
Affine NMT	47.7%	32.9 %

the late reverberation by simply time-shifting the reverberant power spectrum sequence as $\boldsymbol{N}_t = \alpha \boldsymbol{Y}_{t-\Delta}$, where Δ denotes the time-shift amount, which is set at 50 ms, and α is a decay rate that depends on reverberation time. Although [19] requires an estimate of the reverberation time to find an appropriate decay rate, this parameter can be set at an arbitrary value, independent of the reverberation time, in the proposed algorithm. This is because the decay rate is translated into a constant bias of the feature vectors, which is automatically adjusted via the optimization of the affine transformation matrix $\boldsymbol{C}$.

4. *Noise model transfer (using affine transformation):* Based on the late-reverberation power spectra and the reverberant feature vectors, the noise model parameters are optimized using the NMT algorithm described in the previous section.
5. *Vector Taylor series enhancement:* Finally, the enhanced logarithmic mel-frequency spectral feature vectors are calculated with VTS using the reverberant feature vectors and the noise model parameters. The enhanced logarithmic mel-frequency spectra are converted to the mel-frequency cepstral coefficients and their velocity and acceleration coefficients, which are fed into a back-end speech recognizer.

5. EXPERIMENTAL RESULTS

We conducted two reverberant speech recognition experiments to compare the bias and affine transformation models for NMT. In the first experiment, we trained a speaker-independent clean acoustic model using about 250 hours of training data contained in the Wall Street Journal (WSJ) corpus [20]. The acoustic model consisted of phoneme-based left-to-right HMMs with three hidden states per phoneme and ten Gaussians per state. The feature vector used for recognition had 39 dimensions, consisting of 13 static MFCCs (including C0) and their velocity and acceleration coefficients. We used a 20,000-word trigram language model. We created reverberant utterances by convolving clean speech signals with a room impulse response recorded in a meeting room with a reverberation time of 0.78 s and a speaker-to-microphone distance of 2 m. The clean speech signals were taken from the si_et_20 test set of the WSJ corpus, which consists of 333 utterances spoken by eight speakers (four male, four female). The second experiment presumed a more practical situation, where we trained an acoustic model using speech signals with six different reverberation conditions. To generate such multistyle training data, we split the clean training data set into six groups. The speech signals of different groups were convolved with different room impulse responses. The reverberation time of the training impulse responses ranged from 0 (clean) to 1.3 s. The other experimental conditions were unchanged from the first experiment.

Table 1 lists the word error rates (WERs) we obtained under different experimental conditions. The baseline clean speech WER was 10.2%, which is slightly larger than the state-of-the-art WSJ task recognition performance. This results from both the use of C0 instead of the log energy and the use of a classical recognizer with a likelihood maximizing acoustic model and a conventional trigram language model. With the clean acoustic model, the WER sharply increased to 93.5% for reverberant speech. The WER was reduced to 55.9 % when we performed feature enhancement prior to recognition using bias transformation-based NMT. The affine transformation further reduced the WER to 47.7%, indicating a relative WER reduction of 14.7 % in comparison with the bias transformation model. The multistyle training improved the speech recognition performance and achieved a WER of 44.8% for reverberant speech. Feature enhancement preprocessing using bias and affine transformation models reduced the WER to 35.1% and 32.9%, respectively, which means that the affine transformation model achieved a relative WER reduction of 6.27% over the bias transformation model. Note that the multistyle acoustic model used for recognizing the enhanced feature vectors was trained on the feature vectors that were obtained by enhancing the training reverberant utterances, which is sometimes called adaptive training [21, Section 33.7.1]. It is important to note that, even with multistyle training, the affine transformation model improved the speech recognition performance compared with the bias transformation model.

6. CONCLUSION

This paper presented an algorithm for performing NMT using an affine transformation, which has more adjustable parameters and thus is more flexible than the bias transformation model that was previously used. The feature-domain noise model parameter values produced by NMT can be used to enhance corrupted feature vectors. The present method was applied to reverberant speech recognition by regarding late reverberation as nonstationary additive noise. Experimental results for large-vocabulary reverberant speech recognition showed that the affine transformation model eliminated more speech recognition errors than the bias transformation model.

7. APPENDIX

- Conditional pdf of corrupted feature vector

$$p_{\mathrm{y|n,k}}(\boldsymbol{y}_t|\boldsymbol{n}_t,k;\boldsymbol{h}\backslash\boldsymbol{x}',\boldsymbol{n}',\boldsymbol{h}') = f_{\mathrm{N}}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_{t,k}^{\mathrm{y|n}},\mathrm{diag}\left(\boldsymbol{\sigma}_{t,k}^{\mathrm{y|n}}\right)\right)$$
$$\boldsymbol{\mu}_{t,k}^{\mathrm{y|n}} = f(\boldsymbol{x}',\boldsymbol{n}',\boldsymbol{h}') + (1-g(\boldsymbol{x}',\boldsymbol{n}',\boldsymbol{h}'))(\boldsymbol{n}_t-\boldsymbol{n}') + g(\boldsymbol{x}',\boldsymbol{n}',\boldsymbol{h}')(\boldsymbol{\mu}_k^{\mathrm{x}}-\boldsymbol{x}'+\boldsymbol{h}-\boldsymbol{h}')$$
$$\boldsymbol{\sigma}_{t,k}^{\mathrm{y|n}} = g(\boldsymbol{x}',\boldsymbol{n}',\boldsymbol{h}')^2\boldsymbol{\sigma}_k^{\mathrm{x}}$$

- Marginal pdf of corrupted feature vector

$$p_{\mathrm{y|k}}(\boldsymbol{y}_t|k;\theta_t,\boldsymbol{h}) = f_{\mathrm{N}}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_{t,k}^{\mathrm{y}},\mathrm{diag}\left(\boldsymbol{\sigma}_{t,k}^{\mathrm{y}}\right)\right)$$
$$\boldsymbol{\mu}_{t,k}^{\mathrm{y}} = f(\boldsymbol{\mu}_k^{\mathrm{x}},\boldsymbol{\mu}_t^{\mathrm{n}},\boldsymbol{h})$$
$$\boldsymbol{\sigma}_{t,k}^{\mathrm{y}} = g(\boldsymbol{\mu}_k^{\mathrm{x}},\boldsymbol{\mu}_t^{\mathrm{n}},\boldsymbol{h})^2\boldsymbol{\sigma}_k^{\mathrm{x}} + (1-g(\boldsymbol{\mu}_k^{\mathrm{x}},\boldsymbol{\mu}_t^{\mathrm{n}},\boldsymbol{h}))^2\boldsymbol{\sigma}_t^{\mathrm{n}}$$

- Posterior probability of Gaussian index

$$\gamma_{t,k} = \frac{\pi_k p_{\mathrm{y|k}}(\boldsymbol{y}_t|k;\theta_t,\boldsymbol{h})}{\sum_{k'}\pi_{k'}p_{\mathrm{y|k}}(\boldsymbol{y}_t|k';\theta_t,\boldsymbol{h})}$$

- Posterior pdf of additive noise feature vector

$$p_{\mathrm{n|y,k}}(\boldsymbol{n}_t|\boldsymbol{y}_t,k;\theta_t,\boldsymbol{h}) = f_{\mathrm{N}}\left(\boldsymbol{n}_t;\boldsymbol{\mu}_{t,k}^{\mathrm{n|y}},\mathrm{diag}\left(\boldsymbol{\sigma}_{t,k}^{\mathrm{n|y}}\right)\right)$$
$$\boldsymbol{\mu}_{t,k}^{\mathrm{n|y}} = \boldsymbol{\mu}_t^{\mathrm{n}} + \boldsymbol{\xi}_{t,k}^{\mathrm{n}}g(\boldsymbol{\mu}_k^{\mathrm{x}},\boldsymbol{\mu}_t^{\mathrm{n}},\boldsymbol{h})(\boldsymbol{y}_t-\boldsymbol{\mu}_{t,k}^{\mathrm{y}})$$
$$\boldsymbol{\sigma}_{t,k}^{\mathrm{n|y}} = \boldsymbol{\xi}_{t,k}^{\mathrm{n}}\boldsymbol{\sigma}_{t,k}^{\mathrm{y|n}}$$
$$\boldsymbol{\xi}_{t,k}^{\mathrm{n}} = \frac{\boldsymbol{\sigma}_{t,k}^{\mathrm{n}}}{\boldsymbol{\sigma}_{t,k}^{\mathrm{y}}}$$

8. REFERENCES

[1] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environmental-independent speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1996, vol. 2, pp. 733–736.

[2] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks," in *Proc. Eurospeech*, 2001, pp. 221–224.

[3] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 6, pp. 568–580, 2003.

[4] M. Afify and O. Siohan, "Sequential estimation with optimal forgetting for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 19–26, 2004.

[5] S. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath, "Dynamic noise adaptation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 1197–1200.

[6] T. Yoshioka and T. Nakatani, "Noise model transfer: novel approach to robustness against nonstationary noise," *IEEE Trans. Audio, Speech, Language Process.*, 2012, submitted.

[7] T. Yoshioka and T. Nakatani, "Time-varying residual noise feature model estimation for multi-microphone speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4913–4916.

[8] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[9] H. Liao, *Uncertainty decoding for noise robust speech recognition*, Ph.D. thesis, The University of Cambridge, 2007.

[10] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 2, pp. 231–246, 2009.

[11] E. A. P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, Eindhoven University of Technology, 2006.

[12] K. Kumar, B. Raj, R. Singh, and R. Stern, "An iterative least-squares technique for dereverberation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 5488–5491.

[13] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach," 2006, pp. 1133–1136.

[14] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Commun.*, vol. 50, no. 3, pp. 244–263, 2008.

[15] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1676–1691, 2010.

[16] Y.-Q. Wang and M. J. F. Gales, "Improving reverberant VTS for hands-free robust speech recognition," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2011, pp. 113–118.

[17] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1692–1707, 2010.

[18] H. Kuttruff, *Room Acoustics*, Spon Press, fifth edition, 2009.

[19] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United with Acustica*, vol. 87, pp. 359–366, 2001.

[20] D. B. Paul and J. M. Barker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop. Speech, Natural Language*, 1992, pp. 357–362.

[21] J. Droppo and A. Acero, "Environmental robustness," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., pp. 653–679. Springer, 2008.