EXPLOITING VARIANCES IN ROBUST FEATURE EXTRACTION BASED ON A PARAMETRIC MODEL OF SPEECH DISTORTION

Li Deng, Jasha Droppo, and Alex Acero

Microsoft Research, One Microsoft Way, Redmond WA 98052, USA

ABSTRACT

This paper presents a technique that exploits the denoised speech's variance, estimated during the speech feature enhancement process, to improve noise-robust speech recognition. This technique provides an alternative to the Bayesian predictive classification decision rule by carrying out an integration over the feature space instead of over the model-parameter space, offering a much simpler system implementation and lower computational cost. We extend our earlier work [5] by using a new approach, based on a parametric model of speech distortion and thus free from the use of any stereo training data, to statistical feature enhancement, for which a novel algorithm for estimating the variance of the enhanced speech features is developed. Experimental evaluation using the full Aurora2 test data sets demonstrates an 11.4% digit error rate reduction averaged over all noisy and SNR conditions, compared with the best technique we have developed [2] prior to this work that did not exploit the variance information and that required no stereo training data.

1. INTRODUCTION

Effective exploitation of variances or uncertainty is a key essence in statistical pattern recognition. In already successful applications of HMM-based robust speech recognition, uncertainty in the HMM parameter values has been represented by their statistical distributions (e.g., [9, 8]). The motivation of such model-space Bayesian approaches has been the widely varied speech properties due to many possible sources of differences, including speakers and acoustic environments, across and possibly within training and test data. In order to take advantage of the model parameter uncertainty, the decision rule for recognition or decoding has been improved from the conventional MAP rule to the Bayesian predictive classification (BPC) rule [7]. The former, MAP rule has been

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} p(\mathbf{X}|\bar{\Lambda}, \mathbf{W}) P(\mathbf{W}), \quad (1)$$

where $P(\mathbf{W})$ is the prior probability that the speaker utters a word sequence $\mathbf{W}$, and $P(\mathbf{X}|\bar{\Lambda}, \mathbf{W})$ is the probability that the speaker produces the acoustic feature sequence, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ...\mathbf{x}_T]$ when $\mathbf{W}$ is the intended word sequence. Computation of the probability $P(\mathbf{X}|\bar{\Lambda}, \mathbf{W})$ uses deterministic parameters $\bar{\Lambda}$ in the speech model.

When the parameters Λ of the speech model are made random to take account of their uncertainty, the new BPC rule requires integration over all possible parameter values [7]:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \left[\int_{\Lambda \in \Omega} p(\mathbf{X}|\Lambda, \mathbf{W}) p(\Lambda|\phi, \mathbf{W}) d\Lambda \right] P(\mathbf{W}), \quad (2)$$

where ϕ is the (deterministic) hyper-parameters characterizing the distribution of the random model parameters.

An alternative to the model-space characterization of uncertainty such as the above BPC is to represent the uncertainty by integrating over the feature space instead of over the model parameters. The uncertainty in the feature space can be established during a statistical feature enhancement or extraction process. While most of the feature enhancement algorithms developed in the past discard the uncertainty information [3, 4, 2], such side information available from most of these algorithms can be taken advantage of to improve the recognition decision rule. The main advantage of the feature-space treatment of uncertainty over that in the model space is the significantly reduced implementation simplicity and computational cost. More detailed motivations for making use of the feature-space uncertainty, called "uncertainty decoding", can be found in our recent work [5], where positive results were reported based on a specific, stereo-based feature enhancement algorithm (SPLICE [3, 4]) under a matched training and testing condition.

To relax the matched condition required of the SPLICE for effective uncertainty decoding, we in this paper will present a new uncertainty decoding technique based on a statistical enhancement algorithm developed using a parametric model of speech distortion and hence free from any stereo training data. In Section 2, we will introduce this technique. Detailed computation for the variances required by the technique will be presented in Section 3. Comprehensive results obtained using the complete Autora2 task will be reported in Section 4. They demonstrate the effectiveness of feature-space uncertainty decoding for noise-robust speech recognition under the full range of noisy and SNR conditions supplied by the Aurora2 database.

2. NEW DECISION RULE EXPLOITING VARIANCE IN FEATURE EXTRACTION

As discussed above, uncertainty decoding based on the feature-space variance information provides greater simplicity compared with the model-space uncertainty decoding strategy exemplified by the BPC decision rule of Eq.2. The counterpart of the BPC rule in the feature space requires an integration over the uncertainty in the enhanced feature sequence $\hat{\mathbf{X}}$ rather than over that in the model parameter Λ:

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \left[\int_{\hat{\mathbf{X}} \in \Psi} p(\hat{\mathbf{X}}|\bar{\Lambda}, \mathbf{W}) p(\hat{\mathbf{X}}|\theta) d\hat{\mathbf{X}} \right] P(\mathbf{W}), \quad (3)$$

where $\bar{\Lambda}$ is the fixed model parameters (no uncertainty), and θ is the parameters characterizing the distribution, $p(\hat{\mathbf{X}}|\theta)$, of the enhanced speech features computed from a statistical feature extraction algorithm. Note that, unlike the model-domain uncertainty

characterization by $p(\Lambda|\phi, \mathbf{W})$ in Eq.2, $p(\hat{\mathbf{X}}|\theta)$ in Eq.3 can be reasonably assumed to be independent of the word identities $\mathbf{W}$ (and independent of model parameters $\bar{\Lambda}$).

The main motivation for the use of the new decision rule Eq.3 is the acceptance that no noise reduction or feature enhancement algorithm is perfect. Use of an estimated degree of the imperfection according to the distribution $p(\hat{\mathbf{X}}|\theta)$ provides a mechanism to effectively mask some undesirable distortion effects. For example, the frames with a negative instantaneous SNR which are difficult to enhance can be automatically discounted when the variance in $p(\hat{\mathbf{X}}|\theta)$ for these frames is sufficiently large. This mechanism may also effectively extend the HMM uncertainty to cover the gap between true and estimated clean speech features.

For simplicity purposes, we in this paper use the Gaussian assumption to characterize the uncertainty in the enhanced speech features:

$$p(\hat{\mathbf{x}}_t|\theta_t) = \mathcal{N}(\hat{\mathbf{x}}_t; \boldsymbol{\mu}_{\hat{\mathbf{x}}_t}, \boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}), \tag{4}$$

where frame (t)-specific parameters are $\theta_t = [\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}, \boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}]$. Note that the variance parameter $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$ provides a complete characterization of the uncertainty. The Gaussian distributions are assumed independent across frames.

There are hence two key issues concerning the use of the new decision rule Eq.3 that exploits variances in statistical feature extraction or enhancement for improving noise-robust speech recognition. The first issue is: given an estimate of the uncertainty (described by $p(\hat{\mathbf{X}}|\theta)$ for whatever feature enhancement algorithm in use), how to incorporate it into the recognizer's decoding rule? Consider an HMM system with a mixture of Gaussians as the output distribution. Under the Gaussian assumption of Eq.4 for the feature-uncertainty characterization, it can be shown that the integral (i.e., the acoustic score in uncertainty decoding):

$$\int_{\hat{\mathbf{X}} \in \Psi} p(\hat{\mathbf{X}}|\bar{\Lambda}, \mathbf{W}) p(\hat{\mathbf{X}}|\theta) d\hat{\mathbf{X}}$$

in Eq.3 is close to the conventional acoustic score (MAP decoding in Eq.1) when the variance of each Gaussian in the HMM is increased by the amount equal to $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$ on a frame-by-frame basis. This is significantly simpler to implement than the model-space integration in Eq.2.

The second issue is: how to estimate the uncertainty in statistical feature enhancement? We address this issue in the next section in the context of a specific feature enhancement algorithm based on a specific parametric model of speech distortion.

3. COMPUTING UNCERTAINTY BASED ON A PARAMETRIC MODEL OF SPEECH DISTORTION

3.1. Overview of a parametric model of speech distortion

The parametric model of speech distortion, similar to the ones described earlier in [2, 6], is briefly reviewed here as the basis for robust feature extraction from which the uncertainty (i.e., the Gaussian variance $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$ in Eq.4) is computed. Let $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{n}$ be single-frame vectors of log Mel-filter energies for the noisy speech, clean speech, and additive noise, respectively. These quantities can be shown to be governed by the following relationship:

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + \log\left[(\mathbf{1} + e^{\mathbf{n}-\mathbf{x}})[\mathbf{1} + 2\,\boldsymbol{\lambda} e^{\frac{\mathbf{n}-\mathbf{x}}{2}} / (\mathbf{1} + e^{\mathbf{n}-\mathbf{x}})]\right] \\ &\approx \mathbf{x} + \log(\mathbf{1} + e^{\mathbf{n}-\mathbf{x}}) + \boldsymbol{\lambda} / \cosh(\frac{\mathbf{n}-\mathbf{x}}{2}) \end{aligned} \tag{5}$$

where $\boldsymbol{\lambda}$ is the inner product between the clean speech and noise vectors of Mel-filter energies in the linear domain, and the last step of approximation uses the assumption that $\boldsymbol{\lambda} \ll \cosh(\frac{\mathbf{n}-\mathbf{x}}{2})$.

In order to avoid complicated evaluation of the small prediction residual in Eq.5:

$$\mathbf{r} = \boldsymbol{\lambda} / \cosh(\frac{\mathbf{n}-\mathbf{x}}{2}), \tag{6}$$

it is represented by an "ignorance" model as a zero-mean, Gaussian random vector. This thus gives a parametric model of

$$\mathbf{y} = \mathbf{x} + \mathbf{g}(\mathbf{n} - \mathbf{x}) + \mathbf{r}, \tag{7}$$

where $\mathbf{g}(\mathbf{z}) = \log(\mathbf{1} + e^{\mathbf{z}})$, and $\mathbf{r} \sim \mathcal{N}(\mathbf{r}; \mathbf{0}, \boldsymbol{\Psi})$.

The Gaussian assumption for the residual $\mathbf{r}$ in model of Eq.8 allows straightforward computation of the likelihood for the noisy speech vector according to

$$p(\mathbf{y}|\mathbf{x}, \mathbf{n}) = \mathcal{N}(\mathbf{y}; \mathbf{x} + \mathbf{g}(\mathbf{n} - \mathbf{x}), \boldsymbol{\Psi}) \tag{8}$$

3.2. Computing expectations of enhanced speech features

We now discuss the computation of expectations of enhanced speech features as the MMSE estimate of clean speech given the speech distortion model Eq.8. The computation of the MMSE estimate presented here uses a prior clean-speech model for the joint static feature $\mathbf{x}_t$ and delta feature $\Delta\mathbf{x}_t \equiv \mathbf{x}_t - \mathbf{x}_{t-1}$ according to the following Gaussian-mixture distribution:

$$p(\mathbf{x}_t, \Delta\mathbf{x}_t) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x) \mathcal{N}(\Delta\mathbf{x}_t; \boldsymbol{\mu}_m^{\Delta\mathbf{x}}, \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}).$$

For simplicity purposes, the prior model for noise is assumed to be a time-varying Dirac delta function:

$$p(\mathbf{n}_t) = \delta(\mathbf{n}_t - \bar{\mathbf{n}}_t), \tag{9}$$

where $\bar{\mathbf{n}}_t$ is computed by a noise tracking algorithm described in [1] and is assumed to be known in the following description of the iterative MMSE estimation for the clean speech vectors.

Some derivation steps for the MMSE estimate described below have been given in [2]. First, train and fix all parameters in the clean speech model: $c_m, \boldsymbol{\mu}_m^x, \boldsymbol{\mu}_m^{\Delta\mathbf{x}}, \boldsymbol{\Sigma}_m^x$, and $\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}$. Then, compute the noise estimates, $\bar{\mathbf{n}}_t$, and compute the weighting matrices:

$$\begin{aligned} \mathbf{V}_1(m) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi} (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1} (\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}), \\ \mathbf{V}_2(m) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi} (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1} \boldsymbol{\Sigma}_m^x, \\ \mathbf{V}_3(m) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}_m^x, \end{aligned}$$

Next, fix the total number, J, of intra-frame iterations. (Iterations are used to approximate the nonlinear function $\mathbf{g}(\mathbf{n} - \mathbf{x})$ in Eq.7 using Taylor series expansion). For each frame $t = 2, 3, ..., T$ in a noisy utterance $\mathbf{y}_t$, set iteration number $j = 1$, and initialize the clean speech estimator by

$$\hat{\mathbf{x}}_t^{(1)} = arg \max_{\boldsymbol{\mu}_m^x} \mathcal{N}[\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}(\bar{\mathbf{n}}_t - \boldsymbol{\mu}_m^x), \boldsymbol{\Psi}].$$

Then, execute the following steps sequentially over time frames:

- Step 1: Compute

$$\gamma_t^{(j)}(m) = \frac{c_m \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}^{(j)}, \boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})}{\sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}^{(j)}, \boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})},$$

where $\mathbf{g}^{(j)} = \log(\mathbf{1} + e^{\bar{\mathbf{n}}_t - \hat{\mathbf{x}}_t^{(j)}})$.

- Step 2: Update the estimator:

$$\begin{aligned}\hat{\mathbf{x}}_t^{(j+1)} &= \sum_m \gamma_t^{(j)}(m)[\mathbf{V}_1(m)\boldsymbol{\mu}_m^x + \mathbf{V}_2(m)\boldsymbol{\mu}_m^{\Delta\mathbf{x}}] \\ &+ [\sum_m \gamma_t^{(j)}(m)\mathbf{V}_2(m)]\hat{\mathbf{x}}_{t-1}^{(J)} \\ &+ [\sum_m \gamma_t^{(j)}(m)\mathbf{V}_3(m)](\mathbf{y} - \mathbf{g}(\bar{\mathbf{n}} - \hat{\mathbf{x}}_t^{(j)})).\end{aligned}$$

- Step 3: If $j < J$, increment j++, and continue the iteration by returning to Step 1. If $j = J$, then increment t++ and start the algorithm again by re-setting $j = 1$ to process the next time frame until the end of the utterance $t = T$.

The expectation of the enhanced speech feature vector is obtained as the final iteration of the estimate above for each time frame:

$$\boldsymbol{\mu}_{\hat{\mathbf{x}}_t} = \hat{\mathbf{x}}_t^{(J)} \tag{10}$$

3.3. Computing variances of enhanced speech features

Given the expectation for the enhanced speech feature computed above, the variance can now be computed according to

$$\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t} = E[\mathbf{x}_t^2|\mathbf{y}_t] - \boldsymbol{\mu}_{\hat{\mathbf{x}}_t}^2, \tag{11}$$

where

$$E[\mathbf{x}_t^2|\mathbf{y}_t] \approx \frac{\sum_{m=1}^M c_m \overbrace{\int \mathbf{x}_t^2 p(\mathbf{x}_t|m, \hat{\mathbf{x}}_{t-1}) p(\mathbf{y}_t|\mathbf{x}_t, \bar{\mathbf{n}}_t) d\mathbf{x}_t}^{I_m(\mathbf{y}_t)}}{p(\mathbf{y}_t)}. \tag{12}$$

After using the zero-th order Taylor series to approximate the nonlinear function $\mathbf{g}(\mathbf{n}_t - \mathbf{x}_t)$ (contained in $p(\mathbf{y}_t|\mathbf{x}_t, \bar{\mathbf{n}}_t)$; rf. Eq.8) by $\mathbf{g}_0(\bar{\mathbf{n}}_t - \mathbf{x}_0)$, the integral in Eq.12 becomes:

$$\begin{aligned}I_m &= \int \mathbf{x}_t^2 \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\mathcal{N}(\mathbf{y}_t; \mathbf{x}_t + \mathbf{g}_0, \boldsymbol{\Psi}) d\mathbf{x}_t \\ &= \int \mathbf{x}_t^2 \mathcal{N}\left[\mathbf{x}_t; \boldsymbol{\theta}_m(t), (\boldsymbol{\Sigma}_m + \boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m\boldsymbol{\Psi}\right] d\mathbf{x}_t \times N_m(\mathbf{y}_t) \\ &= \left[(\boldsymbol{\Sigma}_m + \boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m\boldsymbol{\Psi} + \boldsymbol{\theta}_m^2\right] \times N_m(\mathbf{y}_t)\end{aligned} \tag{13}$$

where

$$\boldsymbol{\mu}_m = (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}\boldsymbol{\mu}_m^x + (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\boldsymbol{\Sigma}_m^x(\hat{\mathbf{x}}_{t-1} + \boldsymbol{\mu}_m^{\Delta\mathbf{x}}),$$

$$\begin{aligned}\boldsymbol{\Sigma}_m &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\boldsymbol{\Sigma}_m^x\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}, \\ \boldsymbol{\theta}_m(t) &= (\boldsymbol{\Sigma}_m + \boldsymbol{\Psi})^{-1}\left[\boldsymbol{\Psi}\boldsymbol{\mu}_m + \boldsymbol{\Sigma}_m(\mathbf{y}_t - \mathbf{g}_0)\right], \\ N_m(\mathbf{y}_t) &= \mathcal{N}\left[\mathbf{y}_t; \boldsymbol{\mu}_m + \mathbf{g}_0, \boldsymbol{\Sigma}_m + \boldsymbol{\Psi}\right].\end{aligned}$$

Substituting the result of Eq.13 into Eq.12, we obtain

$$E[\mathbf{x}_t^2|\mathbf{y}_t] = \sum_{m=1}^M \gamma_m(\mathbf{y}_t)\left[(\boldsymbol{\Sigma}_m + \boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m\boldsymbol{\Psi} + \boldsymbol{\theta}_m^2(t)\right],$$

where

$$\gamma_m(\mathbf{y}_t) = \frac{c_m N_m(\mathbf{y}_t)}{\sum_{m=1}^M c_m N_m(\mathbf{y}_t)}.$$

Eq.11 then gives the final variance estimate for the (static) enhanced feature. In our implementation, an iterative procedure similar to the computation of the expectations described in Section 3.2 is used to estimate the variance also in order to reduce errors caused by approximating $\mathbf{g}(\mathbf{n} - \mathbf{x})$ by $\mathbf{g}_0(\bar{\mathbf{n}} - \mathbf{x}_0)$.

3.4. Computing variances of temporal differences of the enhanced features

In our implementation, the differentials of the enhanced features, also referred to as the delta or dynamic features, are computed in the same manner as those for the clean speech features:

$$\Delta\hat{\mathbf{x}}_t = \sum_{\tau=-K}^K w_\tau \hat{\mathbf{x}}_{t+\tau}, \;\; \Delta^2\hat{\mathbf{x}}_t = \sum_{\tau=-L}^L v_\tau \Delta\hat{\mathbf{x}}_{t+\tau}, \tag{14}$$

where $K = 3$, $L = 2$, and the weights w_τ and v_τ are fixed. Under the assumption of temporal independence, we can easily determine the variances for these differentials according to

$$\boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t} = \sum_{\tau=-K}^K w_\tau^2 \boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}, \;\; \boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t} = \sum_{\tau=-L}^L v_\tau^2 \boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t}, \tag{15}$$

where $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$ is already computed according to Eq.11.

4. SPEECH RECOGNITION EXPERIMENTS ON THE AURORA2 TASK

We have described the mean and variance estimators (Eqs.10 and 11) that fully characterize the statistical distribution (Eq.4) of the enhanced speech features. Given this distribution, the feature-space uncertainty decoding rule (Eq.3) can be used to perform speech recognition. In the current work, the rule Eq.3 is implemented in the conventional HMM recognizer by adding $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$ to the variances of all Gaussians (about 500 in total for the Aurora2 task using whole-digit units) in the HMMs at each frame t, while using $\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}$ as the observation vector. We have evaluated this new decoding strategy on the Aurora2 database. The task is to recognize strings of connected English digits embedded in several types of artificially created distortion environments with a range of SNRs from 0-20dB. Three sets of digit utterances (sets A, B, and C) are prepared as the test material. The original HMMs used for decoding (before adding the variance estimator $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$) are trained using all clean speech files in the training set of the Aurora2 database. The noise estimate used for computing both the expectations and variances of the enhanced features in the experiments below is based on the iterative stochastic approximation algorithm described in [1].

4.1. Results comparing the uses of uncertainty in different sets of feature streams

Table 1 presents the percent-accurate performance results on all three sets of the Aurora2 test data, averaged over all SNRs from 0 to 20 dB and over four (sets A/B) or two (set C) distortion conditions (each condition and SNR contains 1101 digit strings). Row I gives the baseline results using the conventional MAP rule Eq.1 (i.e., "point" decoding), where the expectations of the enhanced speech feature vectors $\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}$'s computed according to Eq.10 described in Section 3.2 (jointly with $\Delta\hat{\mathbf{x}}_t$ and $\Delta^2\hat{\mathbf{x}}_t$ computed by Eq.14) are used as the observational feature vector sequence $\mathbf{X}$ in Eq.1, and the variances for all feature streams (static and dynamic) are set to zero: $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t} = \boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t} = \boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t} = 0$.

Row II in Table 1 shows the recognizer's performance using the feature-space uncertainty decoding rule Eq.3 where the variance of the static feature stream is computed according to Eq.11 while the variances of the dynamic feature streams are set to zero:

Table 1. Aurora2 performance (percent accurate) exploiting different sets of feature streams. Uncertainty or variances are computed using the estimation formulas described in Section 3.

	set A	set B	set C	Ave.
I: MAP-rule	85.66	86.15	80.40	84.80
II: Static variance only	86.95	87.56	81.62	86.13
III: Static/Δ variances	87.38	87.74	82.44	86.54
IV: Static/Δ/Δ^2 variances	87.34	87.79	82.45	86.54

$\mathbf{\Sigma}_{\Delta\hat{\mathbf{x}}_t} = \mathbf{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t} = 0$. The overall improvement in the recognition accuracy from 84.8% to 86.1% corresponds to an 8.8% digit error rate reduction. The error rate is further reduced, totaling to an 11.4% reduction, when the variances ($\mathbf{\Sigma}_{\Delta\hat{\mathbf{x}}_t}$ and $\mathbf{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t}$) of the dynamic feature streams are estimated by Eq.15 rather than being set to zero (Rows III and IV). But we observed that exploiting the variance of the acceleration feature stream ($\mathbf{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t}$) has not contributed to any performance improvement once the variance of the delta feature stream has been exploited.

4.2. Results on the performance limit of uncertainty decoding

To investigate the upper limit of possible performance improvement by exploiting variances for feature-space uncertainty decoding, we desire to eliminate biases in the variance estimation based on Eqs.11 and 15. To achieve this, we conducted diagnostic experiments where the "true" variances are computed by squaring the differences between the estimated and true clean speech features. The true clean speech features are computed from the clean speech waveforms available from the Aurora2 database, and the estimated clean speech features are computed from Eq.10. The performance results of Table 2 are significantly better than those in Table 1. In particular, we observe that the exploitation of the variances of both the static and the dynamic feature streams cuts the error rate by about half compared with using the variance for the static feature stream only (see the performance difference in Rows I and II of Table 2). In contrast, the corresponding performance difference is much smaller when the estimated variances (as opposed to the true ones) are used. These results suggest that the biases introduced by the variance estimators Eqs.11 and 15 are undesirably large, and that better variance estimators developed in future research will have the potential to drastically improve the recognition performance from those shown in Table 1 towards those in Table 2.

Table 2. Aurora2 performance (percent accurate) using the variances determined by squaring the differences between the estimated and true clean speech features. This eliminates biases in the variance estimation

	set A	set B	set C	Ave.
I: Static variance only	90.31	91.12	84.70	89.51
II: Static/Δ/Δ^2 variances	94.87	95.49	90.75	94.29

5. SUMMARY AND CONCLUSION

The work described in this paper extends our earlier work in speech feature enhancement and noise-robust recognition in two fronts. First, it extends the uncertainty decoding technique [5] by using a new approach, free from the use of any stereo training data, to statistical feature enhancement. Second, it extends the Bayesian technique for speech feature enhancement [2] by exploiting the variance of the enhanced feature via integration over the feature space, leading to the new recognition decision rule. A novel algorithm for estimating the variance, as well as the expectation, of enhanced speech features is developed and described. Experimental evaluation using the full Aurora2 test data sets demonstrates a 11.4% digit error rate reduction, averaged over all noisy and SNR conditions, compared with the best result reported in [2] that did not exploit the variance information.

We also reported the results from a set of diagnostic experiments where the "true" variance is provided to the uncertainty decoding rule so that the gap between the true and the estimated clean speech features is fully covered. More than 50% of the digit errors, committed when the estimated variance is used, have been corrected. This provides a clear direction of our future research on improving the quality of uncertainty estimation within the uncertainty decoding framework presented in this paper.

6. REFERENCES

[1] L. Deng, J. Droppo, and A. Acero. "Recursive estimation of nonstationary noise using a nonlinear model with iterative stochastic approximation," *Proc. ASRU Workshop,* Trento, Italy, Dec. 2001, 4 pages.

[2] L. Deng, J. Droppo, and A. Acero. "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior," *Proc. ICASSP*, Vol.I, Orlando, Florida, May 2002, pp. 829-832.

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

[4] L. Deng, A. Acero, L. Jiang, J. Droppo, and XD Huang. "High-performance robust speech recognition using stereo training data," *Proc. ICASSP*, Vol.1, 2001, pp. 301-304.

[5] J. Droppo, A. Acero, and L. Deng. "Uncertainty decoding with SPLICE for noise robust speech recognition," *Proc. ICASSP*, Vol.I, Orlando, Florida, May 2002, pp. 57-60.

[6] B. Frey, L. Deng, A. Acero, and T. Kristjansson. "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," *Proc. Eurospeech*, 2001, pp. 901-904.

[7] Q. Huo and C. Lee. "A Bayesian predictive approach to robust speech recognition," *IEEE Trans. Speech Audio Proc.*, Vol.8, 2000, pp. 200-204.

[8] H. Jiang and L. Deng. "A robust compensation strategy against extraneous acoustic variations in spontaneous speech recognition," *IEEE Trans. Speech Audio Proc.*, Vol 10, No. 1, 2002, pp. 9-17.

[9] C. Lee, C. Lin, and B. Juang. "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Proc.*, Vol.39, 1991, pp. 806-814.