MODEL-BASED NOISE REDUCTION LEVERAGING FREQUENCY-WISE CONFIDENCE METRIC FOR IN-CAR SPEECH RECOGNITION

Osamu Ichikawa[1], Steven J. Rennie[2], Takashi Fukuda[1], Masafumi Nishimura[1]

[1]IBM Research – Tokyo, Yamato, 242-8502, JAPAN
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
ichikaw@jp.ibm.com, sjrennie@us.ibm.com, {fukuda1, nisimura}@jp.ibm.com,

ABSTRACT

Model-based approaches for noise reduction effectively improve the performance of automatic speech recognition in noisy environments. Most of them use the Minimum Mean Square Estimate (MMSE) criterion for de-noised speech estimates. In general, an observation has speech-dominant bands and noise-dominant bands in the Mel spectral domain. This paper introduces a method to add weight to speech-dominated bands when evaluating the posterior probability of each speech state, as these bands are generally more reliable. To leverage high-resolution information in the Mel domain, we use Local Peak Weight (LPW) as the confidence metric for the degree of speech dominance. This information is also used to regulate the amount of compensation that is applied to each frequency band during feature reconstruction under an integrated probabilistic model. The method produced relative word error rate improvements of up to 33.8% over the baseline MMSE method on an isolated word task with car noise.

Index Terms—Harmonic analysis, speech enhancement, robust speech recognition, model-based noise reduction, missing feature.

1. INTRODUCTION

Various techniques have been studied to improve the performance of automatic speech recognition (ASR) in noisy situations, such as in automobiles. In noise reduction techniques, simple approaches to reducing noise such as spectral subtraction (SS) have had limited success. Their signal recovery is inadequate in very low Signal to Noise Ratio (SNR) situations with ambient noise, such as "Fan high" or "Window open" during high speed driving. For such situations, model-based approaches such as VTS [1], SPLICE [2], and DNA [3] are known to work more effectively. They most often use Minimum Mean Square Estimate (MMSE) criteria for the de-noised speech estimates. These estimates are obtained as weighted sums of the posterior means, where the weights are the posterior probabilities, so the quality of the posterior probability estimation is critically important. If the model is designed in the Mel-log-spectrum domain, we can improve the posterior probability by weighting the bands using frequency-wise confidence metrics, since the use of reliable bands has been widely exploited in previous studies on missing features [4]. This is one of the themes of the work reported here.

The estimates produced by the model-based noise reduction are good approximations in most of the degraded speech cases. However, when the model is mismatched with the current noise condition, distortions derived from the compensation can reduce the accuracy of the ASR. DNA uses Condition Detection [5], which interpolates the observations and the raw estimates to output more observations in mismatched cases. This motivated us to interpolate for each band by using our frequency-wise confidence metrics. If a band is sufficiently clean, then we can use more of those observations. When a band is degraded, we can use more of the compensated values. This is another theme of our work. This paper also shows the two themes can be pursued in one probabilistic model.

The frequency-wise confidence metric can be based upon any indicator of the reliability of the speech band. Many researches on missing feature used local SNR information to identify degraded bands. In this paper, we are interested in using the confidence metric generated from the Local Peak Weight (LPW) [6] to integrate high-resolution information in the Mel band. LPW is a representation of the harmonic structure that is observed in the local peaks at regular intervals in the spectrum domain. In very noisy situations, harmonic structures are often retained only around formant frequencies in vowels. These bands should have more speech power and be more reliable than others. When LPW was originally introduced for speech enhancement, it was called Local Peak Enhancement (LPE) [7]. It relies upon the assumption that the ambient noise does not include harmonic components and can be calculated directly from the observed spectrum on a per-frame basis. Unlike comb filtering, LPE does not require F0 estimation or voiced/unvoiced detection.

Our proposed frequency-wise confidence metric can be applied for any model-based approach as long as it can be modeled in the Mel-log-spectrum domain. For simplicity, this paper uses Segura's MMSE approach [8]. It estimates the MMSE mismatch function using a clean speech GMM and a noise Gaussian, and it subtracts the estimate from the observation in the logarithmic domain. The noise mean and variance should be explicitly given, such as the top 10 frames of each utterance (as used in our experiment in this paper).

2. BASELINE MMSE

We briefly review Segura's MMSE approach, our baseline MMSE system. The output log-power of the band d in the Mel-filter bank at frame t, corresponding to the noisy speech $y_d(t)$ can be written as a function of the energy of the clean speech $x_d(t)$ and the noise $n_d(t)$.

$$y_d(t) = x_d(t) + \log\left(1 + \exp\left(n_d(t) - x_d(t)\right)\right) \quad (1)$$

In vector notation without *t*,

$$\mathbf{y} = \mathbf{x} + \mathbf{g}, \tag{2}$$

where $\mathbf{g}$ is given by the mismatch function $\mathbf{G}$ for each band as

$$g_d = \mathrm{G}_d(\mathbf{x}, \mathbf{n}) = \log(1 + \exp(n_d - x_d)). \tag{3}$$

The clean speech is modeled as a K-Gaussians mixture

$$p(\mathbf{x}) = \sum_k^K \gamma_k \cdot \mathrm{N}(\mathbf{x}; \mu_{x,k}, \Sigma_{x,k}), \tag{4}$$

where γ_k is the *k*-th Gaussian prior probability, with mean $\mu_{x,k}$ and covariance matrix $\Sigma_{x,k}$ (assumed to be diagonal).

In our interpretation, we can model $\mathbf{g}$ as a Gaussian mixture with a 1st order Taylor series,

$$p(\mathbf{g}) = \sum_k^K \gamma_k \cdot \mathrm{N}(\mathbf{g}; \mu_{g,k}, \Sigma_{g,k}), \tag{5}$$

where

$$\mu_{g,k} \cong \log(1 + \exp(\mu_n - \mu_{x,k})) = \mathbf{G}(\mu_{x,k}, \mu_n), \text{ and} \tag{6}$$

$$\Sigma_{g,k} \cong \mathbf{F}(\mu_{x,k}, \mu_n)^2 \cdot (\Sigma_{x,k} + \Sigma_n). \tag{7}$$

$\mathbf{F}$ is an auxiliary function defined for each band as

$$F_d(\mathbf{x}, \mathbf{n}) = (1 + \exp(x_d - n_d))^{-1}. \tag{8}$$

Then the compensated speech $\hat{\mathbf{x}}$ is given with MMSE as

$$\hat{\mathbf{x}} = \mathbf{y} - \int \mathbf{g} \cdot p(\mathbf{g}|\mathbf{y}) d\mathbf{g} \cong \mathbf{y} - \sum_k^K \rho_k(\mathbf{y}) \cdot \mu_{g,k}. \tag{9}$$

The data of $\hat{\mathbf{x}}$ is passed to the backend for recognition.

The posterior probability ρ_k given by $\mathbf{y}$ is

$$\rho_k(\mathbf{y}) = \gamma_k \cdot \mathrm{N}(\mathbf{y}; \mu_{y,k}, \Sigma_{y,k}) \Big/ \sum_{k'} \gamma_{k'} \cdot \mathrm{N}(\mathbf{y}; \mu_{y,k'}, \Sigma_{y,k'}), \tag{10}$$

where

$$\mu_{y,k} \cong \mu_{x,k} + \mathbf{G}(\mu_{x,k}, \mu_n), \text{ and} \tag{11}$$

$$\Sigma_{y,k} \cong \{1 - \mathbf{F}(\mu_{x,k}, \mu_n)^2\} \cdot \Sigma_{x,k} + \mathbf{F}(\mu_{x,k}, \mu_n)^2 \cdot \Sigma_n. \tag{12}$$

2.1 Gain Adaptation

For the baseline MMSE system, we also used a gain adaptation, because the Gaussian modeling in the Mel-log-spectrum domain is dependent on the recording gain. We can regard the observation $\mathbf{y}$ in Section 2 as already pre-processed with the gain adaptation to maximize the total likelihood of the utterance as

$$\mathbf{y} = \mathbf{y}_{in} + q\mathbf{I}, \tag{13}$$

where $\mathbf{y}_{in}$ are the raw observations and q is the scalar bias to adapt the recording gain in the Mel-log-spectrum domain.

3. CONFIDENCE-WEIGHTED MMSE

Assume we have a confidence metric α_d for the *d*-th band. If α_d has a larger value, then the band is probably more reliable as a clue to find the best matching Gaussian. Therefore, we can increase the weight on that band when evaluating the posterior probability using a modified Gaussian with diagonal covariance as

$$\mathrm{N}'(y; \mu_{y,k}, \Sigma_{y,k}) =$$

$$\prod_{d=1}^D \left\{ (2\pi)^{-\frac{1}{2}} \cdot |\Sigma_{y,k,d}|^{-\frac{1}{2}} \cdot \exp\left(-(y_d - \mu_{y,k,d})^2 / 2\Sigma_{y,k,d}\right) \right\}^{\alpha_d} \tag{14}$$

and

$$\rho'_k(\mathbf{y}) = \gamma_k \cdot \mathrm{N}'(\mathbf{y}; \mu_{y,k}, \Sigma_{y,k}) \Big/ \sum_{k'} \gamma_{k'} \cdot \mathrm{N}'(\mathbf{y}; \mu_{y,k'}, \Sigma_{y,k'}), \tag{15}$$

where *D* is the dimension of the Gaussian. It should be noted band *d* is exponentially weighted by α_d in (14). We use (15) instead of (10) as our Confidence Weighted MMSE (CW-MMSE).

This paper uses a confidence metric derived from LPW. It is extracted from an observed signal on a frame-by-frame basis using cut-off DCT operations [6][7]. We obtain Mel-LPW as w_d by processing with the Mel-band-pass filter as

$$w_d = \sum_i v_i \cdot B_{d,i} \Big/ \sum_i B_{d,i}, \tag{16}$$

where $B_{d,i}$ is the *d*-th triangle filter for the *i*-th bin and v_i is the LPW for the *i*-th bin.

As shown in Fig. 1, it is then processed with a sigmoid function and normalized as the confidence metric α_d, where

$$\alpha'_d = 1.0 / (1.0 + \exp(-a \cdot (w_d - 1.0))) \text{ and} \tag{17}$$

$$\alpha_d = \alpha'_d \Big/ \left(\frac{1}{D} \sum_{d'} \alpha'_{d'} \right). \tag{18}$$

The sigmoid function uses the constant a. We used $a = 5$ in our experiments.

As (18) indicates, the average of α is 1. If α is flat and uniformly 1, there is no weighting and N′ becomes a standard Gaussian. We expect α to be close to 1 for unvoiced segments and non-speech segments.

4. CONFIDENCE-WEIGHTED INTERPOLATION

There is a practical choice for the interpolated output between the observation and the compensation, so that we can adjust the weight for each band depending on the confidence metric of that band. The interpolated output $\tilde{x}$ is passed to the backend as

$$\tilde{x}_d = (1.0 - \beta_d) \cdot \hat{x}_d + \beta_d \cdot y_d. \tag{19}$$

We used another confidence metric β varying from 0 to 1 that is also derived from the LPW processed with a sigmoid function,

$$\beta_d = 1.0 / (1.0 + \exp(-a \cdot (w_d - 1.0 - b))), \tag{20}$$

where b is a constant value. We set it to 0.3 in our experiments. If β_d is close to 1, then that band is probably sufficiently clean and needs less compensation. If β_d is close to 0, then the band must have less speech power and be more susceptible to noise and probably requires more compensation.

We call this approach Confidence Weighted Interpolation (CW-INT). This can be combined with CW-MMSE from Section 3 for further improvement.

5. PROBABILISTIC CONFIDENCE-WEIGHTED MMSE

CW-MMSE and CW-INT can be combined in a single probabilistic framework. We suppose the probability of the mismatch vector $\mathbf{g}$ is given by the observation $\mathbf{y}$ and the confidence metric β. They are modeled as

$$p(\mathbf{g}|\mathbf{y}, \beta) = p(\mathbf{g}|\mathbf{y}) \cdot p(\mathbf{g}|\beta), \tag{21}$$

where $p(\mathbf{g}|\beta)$ is a model with a higher probability at g_d=0, as when β_d has a higher value.

Then the distribution of the product probability $p(\mathbf{g}|\mathbf{y},\beta)$ is to be shifted toward g_d =0. If MMSE estimates g_d as being close to zero, then the output becomes close to the observed value, as (9) suggested. This is similar to the behavior of CW-INT.

Our approach models $p(\mathbf{g}|\beta)$ as a Gaussian,

$$p(\mathbf{g}|\beta) = \mathrm{N}(\mathbf{g};0,\psi(\beta)). \quad (22)$$

The variance ψ should have a small value when β is large. There may be various choices for this mapping. We calculated the variance ψ by scaling the variance of the *k*-th Gaussian at the *d*-th band in the speech model as

$$\psi_{k,d} = \Sigma_{x,k,d}\left(\beta_d^{-1} - c\right), \quad (23)$$

where the constant *c* is adjusted between 0 and 1. We set it to 0.5 in our experiments. Then $p(\mathbf{g}|\mathbf{y},\beta)$ can be written as a Gaussian mixture model,

$$p(\mathbf{g}|\mathbf{y},\beta) = \sum_k^K \rho_k''(\mathbf{y})\cdot \mathrm{N}(\mathbf{g};\mu_{g,k},\Sigma_{g,k})\cdot \mathrm{N}(\mathbf{g};0,\psi_k(\beta)) = \sum_k^K \rho_k''(\mathbf{y})\cdot \mathrm{N}(\mathbf{g};\mu_{g,k}'',\Sigma_{g,k}''). \quad (24)$$

The variances and the means are given by

$$\Sigma_{g,k}'' = \left(\Sigma_{g,k}^{-1} + \psi_k^{-1}\right)^{-1} \text{ and} \quad (25)$$

$$\mu_{g,k}'' = \left(\Sigma_{g,k}^{-1}\cdot\mu_{g,k} + \psi_k^{-1}\cdot 0\right)\cdot\Sigma_{g,k}''. \quad (26)$$

The posterior probabilities are given by

$$\rho_k''(\mathbf{y}) = \gamma_k \cdot \mathrm{N}(\mathbf{y};\mu_{y,k}'',\Sigma_{y,k}'') \Big/ \sum_{k'} \gamma_{k'}\cdot \mathrm{N}(\mathbf{y};\mu_{y,k'}'',\Sigma_{y,k'}''), \quad (27)$$

where

$$\Sigma_{y,k}'' = \left(\Sigma_{y,k}^{-1} + \psi_k^{-1}\right)^{-1} \text{ and} \quad (28)$$

$$\mu_{y,k}'' = \left(\Sigma_{y,k}^{-1}\cdot\mu_{y,k} + \psi_k^{-1}\cdot\mu_{x,k}\right)\cdot\Sigma_{y,k}''. \quad (29)$$

Finally, the estimate using MMSE is obtained as

$$\hat{\mathbf{x}} = \mathbf{y} - \int \mathbf{g}\cdot p(\mathbf{g}|\mathbf{y},\beta)d\mathbf{g} \cong \mathbf{y} - \sum_k^K \rho_k''(\mathbf{y})\cdot\mu_{g,k}''. \quad (30)$$

According to (28), the variance used for the posterior probability $\Sigma_{y,k,d}''$ becomes smaller than the original variance $\Sigma_{y,k,d}$ for the *d*-th band when the confidence metric β_d is large. This makes the *d*-th band Gaussian more sensitive. This is similar to the behavior of CW-MMSE.

6. EXPERIMENTS

6.1. Experimental setup

CENSREC-3, a widely used evaluation framework for isolated Japanese word recognition in actual automobile environments, was used in these experiments. This data was collected by the Information Processing Society of Japan (IPSJ), and is often used to evaluate noise reduction algorithms [9]. It has speech data both for training and testing for automatic speech recognition using various kinds of trained acoustic models.

The test data in the database was recorded under 16 environmental conditions using combinations of three vehicle speeds and six kinds of in-car environments. A total of 14,216 utterances spoken by 18 speakers (8 males and 10 females) were recorded at a 16-kHz sampling frequency. The recognition grammar is a list of 50 words.

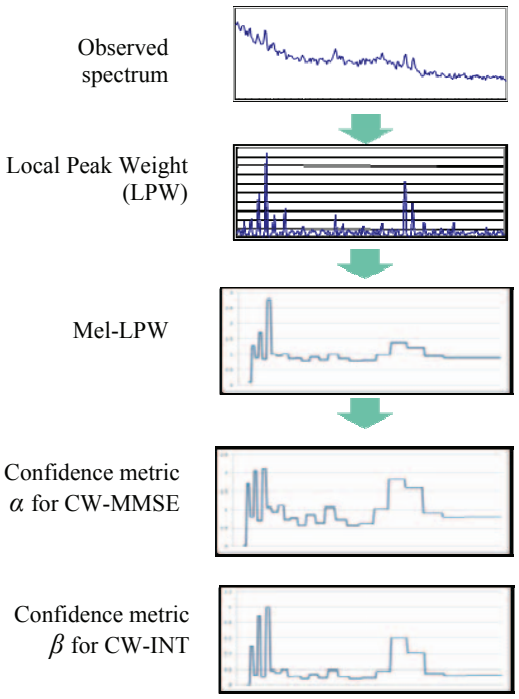

Fig. 1. Process of generating confidence metrics based on LPW information.

For the training data, each driver's voice saying phonetically balanced sentences was recorded under two conditions: while idling and while driving on a city street in a normal in-car environment. A total of 14,050 utterances spoken by 293 drivers (202 males and 91 females) were recorded with a close-talking microphone and a hands-free microphone.

In this experiment, we used only hands-free microphone data for both training and testing. The acoustic models were trained with both idling data and driving data for the front-end processing being tested. This corresponds to Condition 3 as defined in CENSREC-3. For the clean speech GMM, we used idling data recorded with a close-talking microphone. The GMM has 256 Gaussians and it was modeled in a 24-dimensional Mel-log-spectrum domain.

For our evaluation, front-end programs to output various types of features were prepared for the training and the testing, but the backend process to train the acoustic models was unchanged. We used feature vectors with 39 dimensions (12 Mel-cepstrum + C0, with their Δ and ΔΔ) with cepstrum mean normalization (CMN). They were computed from 20-ms frames with a 10-ms shift, using 512 points for the DFT.

6.2. Experimental results

Table 1 shows the resulting word accuracies for various environmental conditions. Since we know the LPW techniques are based on the assumption that the ambient noise does not contain any harmonic component, the audio-on (CD player) cases were excluded from this table. We will discuss the limitations for those cases while explaining Table 2.

Table 1. Results of CENSREC-3 in Condition 3, using 39-dimension feature vectors (12 Mel-cepstrum + C0, with their Δ and ΔΔ) with CMN, in audio-off cases only

CENSREC-3 (Condition 3)		Word Accuracy (%)					
		Baseline	Baseline MMSE	CW-MMSE	CW-INT	CW-MMSE and CW-INT	PCW-MMSE
Idling	Normal	100.0	100.0	99.9	99.8	99.8	100.0
	Hazard on	99.4	97.9	98.0	98.2	98.3	98.2
	Fan low	98.0	98.8	98.8	99.2	99.4	99.1
	Fan high	63.1	81.2	85.2	86.7	88.1	90.2
	Window open	93.1	96.6	96.9	97.6	97.3	97.3
	Average	90.7	94.9	95.8	96.3	96.6	97.0
Low speed	Normal	99.8	98.7	98.9	98.8	99.1	99.2
	Fan low	96.8	97.8	98.0	98.5	97.8	97.9
	Fan high	69.3	84.5	87.4	89.8	89.7	90.8
	Window open	80.8	82.5	85.1	86.7	86.7	88.4
	Average	87.5	91.7	93.0	94.0	94.0	94.7
High speed	Normal	98.1	97.3	97.8	98.3	98.7	98.9
	Fan low	94.8	96.2	96.9	97.2	97.8	98.1
	Fan high	64.8	83.8	85.4	88.7	87.4	89.8
	Window open	49.0	61.5	66.2	67.3	68.8	70.4
	Average	78.8	86.1	87.9	89.1	89.3	90.4
Average		85.2	90.5	91.9	92.8	93.0	93.7

Table 2. Results of CENSREC-3 in Condition 3, in audio-on cases only

CENSREC-3 (Condition 3)		Word Accuracy (%)					
		Baseline	Baseline MMSE	CW-MMSE	CW-INT	CW-MMSE and CW-INT	PCW-MMSE
Idling	Audio on	90.7	88.7	88.1	63.0	73.0	79.5
Low speed	Audio on	93.2	88.7	87.2	72.4	81.6	83.2
High speed	Audio on	91.9	90.4	89.4	81.9	84.8	87.7
Average		91.9	89.3	88.2	72.4	79.8	83.4

The baseline is the evaluation without using any speech enhancement or noise reduction algorithm. The baseline MMSE is Segura's MMSE from Section 2. It has significant gain over the baseline in the "Fan high" and "Window open" cases. The CW-MMSE proposed in Section 3 further reduced the error by 14.4% below the baseline MMSE. The CW-INT proposed in Section 4 reduced the error by 24.4% from the baseline MMSE. It is significant that the band-level interpolation worked very well to improve the performance of the model-based noise reduction approach. The combination of CW-MMSE and CW-INT reduced the error by 26.2% from the baseline MMSE.

PCW-MMSE is the probabilistic confidence-weighting approach proposed in Section 5. It reduced the error by 33.8% from the baseline MMSE. It also outperformed the combination of CW-MMSE and CW-INT. This is an especially promising result, because there is room for improvement in the design of the variance in Equation (23).

Table 2 shows the resulting word accuracies in audio-on cases to evaluate the drawbacks when the noise includes harmonic structures. The baseline MMSE showed a small degradation from the original baseline. Probably, this is caused by the non-stationary nature of the audio noise, because we measured the statistics of the noise only with the top 10 frames of each utterance. All of the proposed methods showed further degradation from the baseline MMSE. This is because LPW has higher values in audio noise. We see driving noise in the high speed case masked the harmonic structure in the audio noise to mitigate the drawbacks.

7. CONCLUSION

In order to improve the performance of model-based noise reduction, we have devised an approach to add weight to the speech-dominant bands by evaluating the posterior probability for the MMSE, favoring the more reliable bands. Another approach uses the interpolated output between the observations and the compensated values, so that we can output more observations instead of the estimated values for those bands, since they must be relatively closer to clean speech. We then combined these two approaches in one probabilistic framework. In order to leverage high-resolution information in the Mel domain, the confidence metric for each band was formulated with harmonic information known as LPW. Experiments showed the proposed approaches successfully outperformed the baseline MMSE.

8. ACKNOWLEDGEMENT

The present study was conducted using the CENSREC-3 database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

9. REFERENCES

[1] P. J. Moreno, B. Raj, and R. M. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proc. of ICASSP*, Vol. II, pp. 733-736, 1996.

[2] J. Droppo, L. Deng and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database." *Proc. of InterSpeech,* pp. 217-220, 2001.

[3] S. J. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath, "Dynamic Noise Adaptation," *Proc. of ICASSP*, Vol. I, pp. 1197-1200, 2006.

[4] B. J. Borgström and A. Alwan, "A Statistical Approach to Mel-Domain Mask Estimation for Missing-Feature ASR," *IEEE Signal Processing Letters*, Vol. 17, pp. 941-944, 2010

[5] S. J. Rennie, P. L. Dognin, and P. Fousek, "Matched-Condition Robust Dynamic Noise Adaptation," *Proc. of ASRU*, pp.137-140, 2011.

[6] O. Ichikawa, T. Fukuda, and M. Nishimura, "DOA estimation with Local-Peak-Weighted CSP," *EURASIP Journal on Advances in Signal Processing*, Article ID 358729, 2010.

[7] O. Ichikawa, T. Fukuda, and M. Nishimura, "Local peak enhancement combined with noise reduction algorithms for robust automatic speech recognition in automobiles," *Proc. of ICASSP*, pp. 4865-4868, 2008.

[8] 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," *Proc. of EuroSpeech,* pp. 221-224, 2001.

[9] M. Fujimoto, et al., "CENSREC-3: Data collection for in-car speech recognition and its common evaluation framework", *Proc. of RWCinME*, pp. 53-60, 2005.