CONSIDERING UNCERTAINTY BY PARTICLE FILTER ENHANCED SPEECH FEATURES IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Matthias Wölfel and Friedrich Faubel

Institut für Theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
wolfel@ira.uka.de

ABSTRACT

The goal of noise compensation techniques is the *perfect* reconstruction of clean features. Unfortunately, the reconstructed features can not be assumed to be perfect. Therefore, to improve performance, the uncertainty of enhanced speech features should be propagated into the hidden Markov model of automatic speech recognition systems.

This paper shows how to jointly estimate the noise and the uncertainty (expressed by the variance) by particle filters in the logarithmic Mel power domain and how to propagate the uncertainty through the front-end into the hidden Markov model. In the experimental section, improvements in word accuracy of a large vocabulary continuous speech recognition system are presented.

Index Terms— uncertainty of enhanced features, dynamic variance compensation, particle filter, noise robust automatic speech recognition,

1. INTRODUCTION

The goal of noise compensation techniques in speech processing and recognition is the *perfect* reconstruction of clean speech features based on an estimate of the noise sequence. One of such techniques which has recently found their way into the speech recognition front-end is the *particle filter* (PF) [1, 2], a.k.a. sequential Monthe Carlo method. The advantage over classical methods such as *spectral subtraction* or *Wiener filters* is the potential to track non-stationary noise. Furthermore, no assumption of the noise distribution has to be taken into account which is one of the mayor drawbacks of *Kalman filters* where the noise distribution has to be assumed to be Gaussian.

Even though PFs overcome some of the former mentioned disadvantages of more traditional and widely used methods, the noise estimate can't be assumed to be perfect. Unfortunately, most of the enhancement techniques presented in the literature have not incorporated the uncertainty of the noise estimate into the *hidden Markov model* (HMM) in particular for large vocabulary tasks. To account for the uncertainty of the noise estimate, Arrowood [3] proposed to replace the point observation of a speech feature by a probability density function. Most of the shown techniques so far, however, rely on stereo data [4], the *signal to noise ratio* [5] (SNR) or are otherwise not jointly estimated by the feature enhancement techniques.

To jointly estimate the mean and variance we propose to use a PF framework in the logarithmic Mel power domain which, for further processing, is propagated through the front-end into the HMM. In the experimental section possible improvements are demonstrated on a large vocabulary continuous speech recognition task coming along with a broad variety of robustness and adaptation methods.

2. HANDLING UNCERTAINTY IN AUTOMATIC SPEECH RECOGNITION

The goal of *automatic speech recognition* (ASR) is to find the most likely word sequence $\hat{\mathbf{W}}$ among all possible word sequences $\mathcal{W}$ given an acoustic observation sequence $\mathbf{x} = [x_1, x_2, \ldots]$. With Bayes' rule we can write this equation as

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in\mathcal{W}}{\operatorname{argmax}}\, p(\mathbf{x}|\Lambda, \mathbf{W})P(\mathbf{W}) \tag{1}$$

where $P(\mathbf{W})$ is the prior probability that the word sequence was uttered, and $p(\mathbf{x}|\Lambda, \mathbf{W})$ is the probability of the acoustic observation sequence $\mathbf{x}$ given the word sequence $\mathbf{W}$ and the acoustic model parameter Λ.

To account for uncertainty in the acoustic observation sequence, (1) has to be extended as follows [4]:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in\mathcal{W}}{\operatorname{argmax}} \left(\int_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}|\Lambda, \mathbf{W}) \cdot p(\mathbf{x}|\theta) d\mathbf{x} \right) P(\mathbf{W}) \tag{2}$$

where $p(\mathbf{x}|\theta)$ represents the distribution of the uncertain acoustic observation sequence and $\mathcal{X}$ represents all possible values of the acoustic observation under the assumption that the uncertainty in the acoustic observation is due to additive noise (in particular its estimate) and therefore independent of the word identities $\mathbf{W}$ and the model parameter Λ.

Considering a HMM with Gaussians as the output distribution

$$p(x|\Lambda_s) = \mathcal{N}(x; \mu_s, \Sigma_s)$$

at state s and denoting the estimated noise mean and estimation error (modeled as a Gaussian distribution where the variance parameter provides a complete characterization of the uncertainty) as

$$p(x|\theta) = \mathcal{N}(x; x - \mu_n, \Sigma_n)$$

we can solve the integral in (2) by the well known equality

$$\int \mathcal{N}(x; \mu_1, \Sigma_1) \cdot \mathcal{N}(x; \mu_2, \Sigma_2) dx = \mathcal{N}(\mu_1; \mu_2, \Sigma_1 + \Sigma_2)$$

as

$$\begin{aligned} \int_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}|\Lambda) \cdot p(\mathbf{x}|\theta) d\mathbf{x} &= \int_{\mathbf{x}} p(\mathbf{x}|\Lambda_s) \cdot p(\mathbf{x}|\theta) d\mathbf{x} \\ &= \int_{\mathbf{x}} \mathcal{N}(\mathbf{x}; \mu_s, \mathbf{\Sigma}_s) \cdot \mathcal{N}(\mathbf{x}; \mathbf{x} - \mu_n, \mathbf{\Sigma}_n) d\mathbf{x} \\ &= \mathcal{N}(\mathbf{x} - \mu_n; \mu_s, \mathbf{\Sigma}_s + \mathbf{\Sigma}_n) \end{aligned} \quad (3)$$

From (3) it follows that the uncleaned feature x is cleaned by subtracting the mean noise estimate μ_n and furthermore, that the Gaussian variance in the HMM process is dynamically compensated by enlarging the acoustic model variance $\mathbf{\Sigma}_s$ associated with clean speech by the variance $\mathbf{\Sigma}_n$ associated with the uncertainty of the noise estimate.

3. PARTICLE FILTER BASED NOISE MEAN AND VARIANCE ESTIMATION

To our best knowledge, Singh and Raj [6] were the first to track the noise sequence that corrupts the speech signal by PFs in the context of speech feature enhancement for speech recognition. The tracked noise sequence is then used to derive an estimate of the clean speech features. An ideal filter would give a perfect estimate of the noise causing the distortion in the representation space, in our case the logarithmic Mel power domain. A real filter, however, can't deliver a perfect estimate. By a small extension to Singh et al. original proposal we are able to derive the uncertainty of the noise estimate modeled as the noise variance.

To model the evolution of noise spectra a 1st-order autoregressive process is used

$$n_t = A \cdot n_{t-1} + \varepsilon_t$$

where A is the transition matrix that is learned either for a specific type of noise or on the silence frames given by voice activity detection, and n_t denotes the noise spectrum at time t. The ε_t terms are considered to be i.i.d. zero mean Gaussian, i.e. $\varepsilon_t \sim \mathcal{N}(0; 0, \Sigma_{\text{noise}})$. Therefore, the noise transition probability $p(n_{t+1}|n_t)$ can be written as

$$p(n_{t+1}|n_t) = \mathcal{N}(n_{t+1}; A \cdot n_t, \Sigma_{\text{noise}}) \quad (4)$$

The clean speech spectra x is modeled in the logarithmic Mel power domain as a Gaussian mixture model

$$p_x(x) = \sum_{k=1}^{K} c_k \mathcal{N}(x; \mu_k, \Sigma_k)$$

trained on speech frames only. Then, using the relationship

$$x_t = y_t + \log(1 - e^{n_t - y_t})$$

between corrupted speech spectra y_t, n_t and x_t, the likelihood

$$l(n_t^{(j)}; y_t) = p(y_t|n_t^{(j)})$$

of a noise hypothesis $n_t^{(j)}$ can be evaluated as

$$p(y_t|n_t^{(j)}) = \frac{p_x(y_t + \log(1 - e^{n_i^{(j)} - y_t}))}{\prod_{i=1}^{d} \left|1 - e^{\hat{n}_{t,i}^{(j)} - y_{t,i}}\right|} \quad (5)$$

If the noise hypothesis $n_t^{(j)}$ exceeds y_t in just one spectral bin, what can happen because noise and speech are not strictly additive in the log domain, the likelihood $p(y_t|n_t^{(j)})$ can't be evaluated and set $p(y_t|n_t^{(j)}) = 0$.

With the above, the particle filter for speech feature enhancement under uncertainty can be outlined as follows:

1. **Sampling**
   For $t = 0$, the noise hypotheses (particles) $n_0^{(j)}$ are drawn from the prior noise density $p(n_0)$. Otherwise, $n_t^{(j)}$ are sampled from the noise transition probability $p(n_t|\bar{n}_{t-1})$ (4).

2. **Calculating the normalized importance weights**
   The importance weight (likelihood) of each noise hypothesis $n_t^{(j)}$ is evaluated if

   $$n_{t,i}^{(j)} < y_{t,i} \; \forall \text{ spectral bins } i$$

   according to (5), otherwise $p(y_t|n_t^{(j)})$ is set to zero. The normalized importance weights are calculated as

   $$\tilde{\omega}_t^{(j)} = \frac{p(y_t|n_t^{(j)})}{\sum_{m=1}^{N} p(y_t|n_t^{(m)})}$$

3. **Noise Mean and Variance Estimation**
   In order to estimate the mean value of the noise we have to sum up the noise estimates according to their weights

   $$\hat{\mu}_{n,t} = \sum_{j=1}^{N} \tilde{\omega}_t^{(j)} \cdot \hat{n}_t^{(j)}$$

   Similar the variance estimate — the uncertainty — can be estimated according to their weights

   $$\hat{\mathbf{\Sigma}}_{n,t} = \tilde{\omega}_t^{(j)} \cdot \left(\hat{n}_t^{(j)} - \hat{\mu}_{n,t}\right) \cdot \left(\hat{n}_t^{(j)} - \hat{\mu}_{n,t}\right)^T$$

4. **Importance resampling**
   The normalized weights are used to resample among the noise hypotheses $n_t^{(1\cdots N)} \rightarrow \bar{n}_t^{(1\cdots N)}$. This can be regarded as a pruning step where likely hypotheses are multiplied, unlikely ones are removed from the population.

Those steps are repeated with $t \mapsto (t+1)$ until all time-frames of the speech data are processed.

More detail on how to infer clean speech, either by a vector tailor series or a direct approach, can be found in [7]. An extensive discussion on PF in ASR including the handling of PF divergence and coupling of the PF by feedback from the ASR system can be found in [8].

4. PROPAGATING UNCERTAINTY THROUGH THE FRONT-END

In order to make use of the uncertainty in the ASR system we have to propagate the uncertainty through the front-end.

To transform the logarithmic Mel power domain into the cepstral domain we have to apply a discrete cosine transformation which can be represented by a multiplication with the matrix $\mathbf{A}_{\mathrm{DCT}}$:

$$\mathbf{x}^{(c)} = \mathbf{A}_{\mathrm{DCT}} \cdot \mathbf{x}^{(l)}$$

To propagate the variance into the cepstral domain we have to multiply the variance matrix with $\mathbf{A}_{\mathrm{DCT}}$ as follows

$$\hat{\mathbf{\Sigma}}_n^{(c)} = \mathbf{A}_{\mathrm{DCT}} \cdot \hat{\mathbf{\Sigma}}_n^{(l)} \cdot \mathbf{A}_{\mathrm{DCT}}^T$$

Given the mean $\mu_x^{(c)}$ and the variance $\mathbf{\Sigma}_x^{(c)}$

$$\mu_x^{(c)} = \frac{1}{T}\sum_t^T \mathbf{x}_t^{(c)}$$

$$\mathbf{\Sigma}_x^{(c)} = \frac{1}{T}\sum_t^T \left(\mathbf{x}_t^{(c)} - \mu_x^{(c)}\right) \cdot \left(\mathbf{x}_k^{(c)} - \mu_x^{(c)}\right)^T$$

we can normalize the values in the cepstral domain by

$$\mathbf{x}_t^{(n)} = \left(\mathbf{x}^{(c)} - \mu_x^{(c)}\right) \div \mathrm{diag}\left(\mathbf{\Sigma}_x^{(c)}\right)$$

where $\div$ stands for component wise division. Similar, the noise variance in the cepstral domain can be normalized by

$$\hat{\mathbf{\Sigma}}_{n,t}^{(n)} = \left(\hat{\mathbf{\Sigma}}_n^{(c)}\right) \div \left(\mathbf{\Sigma}_x^{(c)}\right)^2$$

We then reduce the dimension of 15 consecutive frames by multiplying with the *linear discriminant analysis* (LDA) matrix

$$\mathbf{x}_t^{(s)} = \mathbf{A}_{\mathrm{LDA}} \cdot \mathrm{consecutive}\left(\mathbf{x}_t^{(n)}\right)$$

For the noise variance we get:

$$\hat{\mathbf{\Sigma}}^{(s)} = \mathbf{A}_{\mathrm{LDA}} \cdot \mathrm{consecutive}\left(\hat{\mathbf{\Sigma}}_{n,t}^{(n)}\right) \cdot \mathbf{A}_{\mathrm{LDA}}^T$$

5. SPEECH RECOGNITION EXPERIMENTS

In order to evaluate the improvements by the propagation of the PF uncertainty into the HMM under realistic conditions, we have chosen approximately 45 minutes of lecture speech, taken from the Rich Transcription 2005 Spring Meeting Recognition Evaluation [9], which presents significant challenges to both modeling components used in ASR, namely the language and the acoustic models. To perform experiments on different SNRs we have artificially added, in the time domain, dynamic noise with a broad variety of sounds coming from a truck, slamming containers, distant voices, and shouts [10].

Speech recognition experiments have been performed using the *Janus Recognition Toolkit* (JRTk). We chose to replace the widely used *Mel frequency cestral coefficients* (MFCC)s by warped *minimum variance distortionless response* (MVDR), of model order 60, cepstral coefficients [11]. They have been demonstrated to perform better, on the given data and different SNRs, in comparison to the MFCCs with and without PF [12]. The advantages of the warped MVDR approach over the Fourier transformation are an increase in resolution in low frequency regions, and the dissimilar modeling of spectral peaks and valleys to improve noise robustness, as noise is present mainly in low energy regions. The final features of 42 dimensions were obtained by a LDA transformation on 15 consecutive frames of 13 cepstral mean and variance normalized features. The LDA transformation was followed by a global STC transform [13]. The acoustic training material, approximately 100 hours, used for the experiments reported here was taken from the ICSI, NIST, and CMU meeting corpora, as well as the *Translanguage English Database* (TED) corpus resulting in 3,500 context dependent codebooks with up to 64 Gaussians with diagonal covariances each. The 3-gram language model contained approximately 23,000 words with a perplexity of 125.

Note that the consideration of uncertainty in the HMM is reducing the acoustic score values, since

$$\frac{\sqrt{(\mathbf{x}-\mu_s)^T\left(\mathbf{\Sigma}_s+\mathbf{\Sigma}_n\right)^{-1}(\mathbf{x}-\mu_s)}}{\sqrt{(\mathbf{x}-\mu_s)^T\left(\mathbf{\Sigma}_s\right)^{-1}(\mathbf{x}-\mu_s)}} = \frac{\mathbf{\Sigma}_s}{\mathbf{\Sigma}_s+\mathbf{\Sigma}_n} < 1.0$$

Therefore, for a fair comparison, we have compensated for the offset of the average acoustic score value in the uncertainty case and furthermore worked with wide open beams to not suffer from different pruning depth of the search tree.

Table 1 shows *word errors rates* (WER)s, relative word error reduction and relative uncertainty gain defined as

$$\mathrm{RUG} = \frac{\mathrm{WER}_{\mathrm{uncleaned}} - \mathrm{WER}_{\mathrm{mean\&variance}}}{\mathrm{WER}_{\mathrm{uncleaned}} - \mathrm{WER}_{\mathrm{mean}}} - 1.0$$

For all investigated SNRs we see a clear improvement by the enhanced features on the unadapted as well as on the adapted

Speech Input	Close Talking Speech		SNR 15 dB		SNR 10 dB		SNR 5 dB		SNR 0 dB	
Pass	unadp.	adp.	unadp.	adp.	unadp.	adp.	unadp.	adp.	unadp.	adp.
Compensation	Word Error Rate									
None	31.0%	25.4%	33.3%	27.8%	37.2%	31.3%	45.5%	32.6%	57.8%	42.0%
Mean	—	—	31.6%	27.1%	36.7%	30.2%	43.4%	31.3%	55.1%	39.2%
Mean & Variance	—	—	31.4%	26.9%	36.4%	30.0%	42.9%	31.0%	54.7%	39.0%
Compensation	Relative Word Error Reduction									
Mean	—	—	5.1%	2.5%	1.3%	3.5%	4.6%	4.0%	4.7%	6.7%
Mean & Variance	—	—	5.7%	3.2%	2.2%	4.2%	5.7%	4.9%	5.4%	7.1%
	Relative Uncertainty Gain									
	—	—	11.8%	28.6%	60.0%	18.2%	23.8%	23.1%	14.8%	7.1%

Table 1. Comparison of word error rates, relative word error reduction and relative uncertainty gain for different compensation techniques and *signal to noise ratios* (SNR)s.

passes. The adapted passes have been adapted on the current hypothesis of the unadapted passes. The acoustic models have been adapted by *maximum likelihood linear regression* (MLLR) [14], the features have been adapted by vocal track length normalization and constrained MLLR [13].

Considering the uncertainty of the enhanced features is leading to further improvements in all cases over the results obtained by the PF. In this case the average relative word error reduction is 4.9%, which is an average relative uncertainty gain of 19.2%.

6. CONCLUSIONS

We have presented improvements in accuracy of a large vocabulary continuous speech recognition system by jointly estimating the noise and the uncertainty in the logarithmic Mel power domain by particle filters and its propagation into the hidden Markov model. In the future we want to investigate more reliable uncertainty estimates and its propagation through the front-end.

7. ACKNOWLEDGMENT

The work presented here was partly funded by the *European Union* (EU) under the project CHIL (Grant number IST-506909).

8. REFERENCES

[1] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," *Proc. of ICASSP*, 2004.

[2] M. Fujimoto and S. Nakamura, "Particle filter based non-stationary noise tracking for robust speech feature enhancement," *Proc. of ICASSP*, 2005.

[3] J.A. Arrowood, *Using observation uncertainty for robust speech recognition*, Ph.D. Thesis. Georgia Institute of Technology, 2003.

[4] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," *Proc. of ICASSP*, 2002.

[5] M.C. Benitez, J.C. Segura, A. dela Torre, J. Ramirez, and A. Rubio, "Including uncertainty of speech observations in robust speech recognition," *Proc. of ICSLP*, 2004.

[6] B. Raj and R. Singh, "Tracking noise via dynamical systems with a continuum of states," *Proc. of ICASSP*, 2003.

[7] F. Faubel and M. Wölfel, "Overcoming the vector tailor series approximation in speech feature enhancement – a particle filter approach," *Proc. of ICASSP*, 2007.

[8] F. Faubel, *Speech Feature Enhancement for Speech Recognition by Sequential Monte Carlo Methods*, Diploma Thesis. Universität Karlsruhe (TH), Germany, Aug. 2006.

[9] NIST, "Rich transcription 2005 spring meeting recognition evaluation," *www.nist.gov/speech/tests/rt/rt2005/spring*.

[10] The Freesound Project, "garbage.coll.serv.ds70p.mp3," *freesound.iua.upf.edu/samplesViewSingle.php?id=6986*.

[11] M. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[12] F. Faubel and M. Wölfel, "Coupling particle filters with automatic speech recognition for speech feature enhancement," *Proc. of Interspeech*, Sep. 2006.

[13] M.J.F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[14] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, pp. 171–185, 1995.