SPEECH RECOGNITION WITH SPEECH DENSITY ESTIMATION BY THE DIRICHLET PROCESS MIXTURE

Kenko OTA[†,‡], Emmanuel DUFLOS[†], Philippe VANHEEGHE[†] and Masuzo YANAGIDA[‡]

[†]LAGIS (UMR CNRS 8146), Ecole Centrale de Lille
BP48, Cité Scientifique, 59651, Villeneuve d'Ascq, France
[‡]Doshisha University, Dept. of Engineering
1-3, Tatara-Miyakodani, Kyotanabe, Kyoto, 610-0321, Japan

ABSTRACT

This paper shows a method for the modeling of speech signal distributions based on Dirichlet Process Mixtures (DPM) and the estimation of noise sequences based on particle filtering. In real situations, the speech recognition rate degrades miserably because of the effect of environmental noises, reflected waves and so on. To improve the speech recognition rate, a technique for the estimation of noise sequences is necessary. In this paper, the distribution of the clean speech is modeled using the DPM instead of the traditional model, which is a Gaussian Mixture Model (GMM). Speech signal sequences are generated according to the mean and covariance generated from the DPM. Then, noise signal sequences are estimated with a particle filter. The proposed method using Extended Kalman Filter (EKF) can improve the speech recognition rate significantly in the low SNR region. Applying Unscented Kalman Filter (UKF), better results can be obtained in also the high SNR.

Index Terms— Kalman filtering, Signal processing, Speech enhancement, Speech recognition, Stochastic processes

1. INTRODUCTION

This paper proposes a technique for the estimation of noise and speech sequences without developing the GMM. Instead of the GMM, the speech distribution is modeled using a DPM [2]. The Dirichlet Process (DP) [3] is a non-parametric probability distribution over the space of all possible distributions. The DP is used as the prior of the DPM. The DP can be considered as the probability distribution for the probability distribution of mixture components. The DP is a generative model for infinite distribution. So, DPM allows us to mix the infinite probability distribution. By using DPM in the estimation process of the clean speech distribution, it is expected to estimate this distribution more flexibly.

There are several researches on the nonparametric density estimation using DPM [4], [5]. Caron *et al.* [6] applied the DPM to the density estimation in the context of dynamic models. Caron *et al.* can achieve the improvement of the performance of standard algorithms when the noise pdfs are unknown. Hence, in case where the clean speech distributions are unknown, we also expect to get better result than the standard algorithms.

This paper is organized in the following four sections: the section 2 describes the proposed method, the section 3 shows the evaluation of the proposed method on the speech recognition and the section 4 concludes this paper.

2. PROPOSED METHOD

We propose the modeling of the clean speech using DPM instead of GMM. By introducing DPM, we expect more flexible estimation of clean speech. Because DPM allows us to mix infinite probability distribution. Moreover, DPM can adapt automatically the number of gaussian laws needed. If we want to mix other laws than gaussian, it is also possible.

2.1. Dynamic model for the proposed method

We, as well as Fujimoto *et al.*, employed the dynamic model proposed by Segura *et al.* for each particle as follows [7]:

$$x_t = s_t + \log(1 + \exp(n_t - s_t)) + v_t \quad (1)$$

$$n_t = n_{t-1} + w_{t-1} \quad (2)$$

$$v_t \sim \mathcal{N}(0, \Sigma_s), w_t \sim \mathcal{N}(0, \Sigma_w)$$

where, t is a frame index, x_t is a observed signal, s_t is a clean speech, n_t is a noise signal, $\mathcal{N}(\cdot)$ is a gaussian distribution, v_t and w_t are independent.

2.2. Dirichlet Processes

Ferguson *et al.* [3] defined two properties for the adequate *a priori* distribution.

1. The support of the prior distribution should be large.
2. Posterior distribution given a sample of observation from the true probability distribution should be manageable analytically.

In [3], the authors introduced the DP as a probability measure on the space of probability measures, which satisfies the above properties.

Many probability distributions can be obtained using urn models. The urn model that corresponds to the Dirichlet distribution is the Polya urn model [8]. Here, a probability distribution $\mathbb{G}$ is drawn from $DP(\mathbb{G}_0, \alpha)$ where a probability measure $\mathbb{G}_0$ is defined on a measurable space $(\Omega, \mathcal{A})$, α is a positive real number called scale factor. Let θ_t be a random sample from $\mathbb{G}$. Blackwell *et al.* [8] showed that the predictive distribution is given by the Polya urn model as follows

$$\theta_{t+1}|\theta_t \sim \frac{\alpha}{\alpha+t}\mathbb{G}_0 + \frac{1}{\alpha+t}\Sigma_{j=1}^{t}\delta(\theta-\theta_j).$$

where $\delta(\cdot)$ is the delac delta function.

2.3. Dirichlet Process Mixture

It is possible to reformulate the density estimation problem using the following hierarchical model known as DPM [6]:

$$\mathbb{G} \sim DP(\mathbb{G}_0, \alpha), \;\; \theta_t \sim \mathbb{G}, \;\; s_t \sim f(\cdot|\theta_t) \tag{3}$$

where the RPM (Random Probability Measure) $\mathbb{G}$ is the mixture distribution distributed according to $DP(\mathbb{G}_0, \alpha)$. The latent variables θ_t are distributed according to $\mathbb{G}$. $f(\cdot|\theta_t)$ is a mixed probability density function. The following flexible model is adopted for the unknown distribution $F(s) = \int_\Theta f(s|\theta)d\mathbb{G}(\theta)$ with $\theta \in \Theta$.

2.4. Estimation of speech signal distribution with the Dirichlet process mixture

In the bayesian framework, our problem of estimating a noise sequence and a clean speech sequence, is equivalent to the determination of the probability $p(n_{0:t}, s_{1:t}|x_{1:t})$. A clean speech s_t is supposed to be distributed according to a DPM of base mixed distribution $\mathcal{N}(\mu_t, \Sigma_t)$ and scale parameter α [6]. Instead of developing an accurate GMM, we introduce the estimation of clean speech signal distribution with the DPM which will adapt automatically the number of Gaussian laws to use for the modeling of the clean speech. The problem is now to determine the probability $p(n_{0:t}, \theta_{1:t} \mid x_{1:t})$, decomposed as $p(n_{0:t}, \theta_{1:t} \mid x_{1:t}) = p(n_{0:t} \mid \theta_{1:t}, x_{1:t})p(\theta_{1:t} \mid x_{1:t})$ where, θ_t consists of the mean vector μ_t and covariance matrix Σ_t of clean speech signal. θ_t and a clean speech are drawn from the hierarchical model shown in eq. (3).

$\mathbb{G}_0$ denotes a Normal-inverse Wishart distribution which is usually used when θ_t are a mean μ and a covariance Σ of gaussian law: $\mathbb{G}_0 = \mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$ with $\mu_0, \kappa_0, \nu_0, \Lambda_0$ the hyperparameters of the Normal-inverse Wishart [9].

As $p(n_{0:t} \mid \theta_{1:t}, x_{1:t})$ can be computed using the EKF defined by Fujimoto *et al.* [1] as well as the UKF, we only need to estimate $p(\theta_{1:t} \mid x_{1:t})$ using a particle method. At time t, it follows that $p(n_t, \theta_{1:t} \mid x_{1:t})$ is approximated through a set of J particles by the following empirical distribution

$$P_N(n_t, \theta_{1:t} \mid x_{1:t}) = \Sigma_{j=1}^{J}\tilde{\omega}_t^{(j)} p(n_t \mid \theta_{1:t}^{(j)}, x_{1:t})$$

with $p(n_t \mid \theta_{1:t}^{(j)}, x_{1:t}) \simeq \mathcal{N}(\hat{n}_{t|t}(\theta_{1:t}^{(j)}), \Sigma_{n_{t|t}}^{(j)}(\theta_{1:t}^{(j)}))$. The parameters $\hat{n}_{t|t}(\theta_{1:t}^{(j)})$ and $\Sigma_{n_{t|t}}^{(j)}(\theta_{1:t}^{(j)})$ are computed recursively for each particle j using the EKF. The posterior $p(\theta_{1:t}^{(j)} \mid x_{1:t})$ is proportional to $p(\theta_{1:t-1}^{(j)} \mid x_{1:t-1})$ as follows:

$$\begin{aligned} & p(\theta_{1:t}^{(j)} \mid x_{1:t}) \\ \propto \;\; & p(\theta_{1:t-1}^{(j)} \mid x_{1:t-1}) p(x_t \mid \theta_{1:t}^{(j)}, x_{1:t-1}) p(\theta_t^{(j)} \mid \theta_{1:t-1}^{(j)}) \end{aligned}$$

where

$$\begin{aligned} p(x_t \mid \theta_{1:t}^{(j)}, x_{1:t-1}) &= p(x_t \mid \theta_t^{(j)}, \theta_{1:t-1}^{(j)}, x_{1:t-1}) \\ &= \mathcal{N}(\hat{x}_t(\theta_{1:t}^{(j)}), \hat{\Sigma}_x(\theta_{1:t}^{(j)})) \end{aligned}$$

and

$$\begin{aligned} \hat{x}_t(\theta_{1:t}^{(j)}) &= s_t^{(j)} + \log(I + \exp(n_t^{(j)} - s_t^{(j)})) \\ \hat{\Sigma}_x(\theta_{1:t}^{(j)}) &= G_t^{(j)} \Sigma_{n_t}^{(j)} G_t^{(j)T} + \Sigma_{s,t} \\ G_t^{(j)} &= \frac{\partial}{\partial n_t^{(j)}} \left\{ s_t^{(j)} + \log(1 + \exp(n_t^{(j)} - s_t^{(j)})) \right\} \\ s_t^{(j)} &\sim \mathcal{N}(\mu_t^{(j)}, \Sigma_t^{(j)}) \end{aligned}$$

Finally, sample weights are calculated using these estimates.

$$\tilde{\omega}_t^{(j)} \propto \omega_{t-1}^{(j)} \mathcal{N}(\hat{x}_t(\theta_{1:t}^{(j)}), \hat{\Sigma}_x(\theta_{1:t}^{(j)}))$$

because we chose the importance distribution as follows:

$$q(\theta_t|\theta_{1:t-1}^{(j)}, x_{1:t}) = p(\theta_t|\theta_{1:t-1}^{(j)}).$$

$p(\theta_t^{(j)} \mid \theta_{t-1}^{(j)})$ is determined using the polya urn representation [6].

2.5. Detection of speech/non-speech frame

In the high SNR region, there was the possibility that the noise tracking performance by the proposed method degrade [10]. In this paper, we introduce detection of speech/non-speech frame into the proposed method. Detection is performed based on the distance defined as follows:

$$\begin{aligned} d_{s_t} &= (x_t - (\hat{s}_t + \log(1 + \exp(\hat{n}_t - \hat{s}_t))))^2 \\ d_{n_t} &= (x_t - \hat{n}_t)^2 \\ \Delta d_t &= d_{s_t} - d_{n_t} \end{aligned}$$

where $\hat{s}_t$ and $\hat{n}_t$ are the estimated clean speech and noise signal. If Δd_t is larger than a threshold obtained from the average of Δd_t over first 5 frames [1], the current frame is considered as the speech frame and modified as follows:

$$\hat{s}_t = \hat{s}_t + \xi_s\sqrt{d_{s_t}}, \;\; \hat{n}_t = \hat{n}_t - \xi_n\sqrt{d_{s_t}}$$

[1] We assume first 5 frames are the noise frames.

where ξ_s and ξ_n are determined by the wiener coefficient. In the reverse case, the signs of above equations are inverted.

The proposed method can finally be represented as the following algorithm.

$j = 1, \cdots, J$
  $n_0^{(j)} \sim \mathcal{N}(\mu_N, \Sigma_N) \quad \omega_t^{(j)} = \frac{1}{J}$
end
$t = 1, \cdots, T$
  calculate μ_0, Λ_0
  $j = 1, \cdots, J$
    if $t == 1$ $\quad \theta_t^{(j)} \sim \mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$
    else $\quad \theta_t^{(j)} \sim p(\theta_t^{(j)} | \theta_{t-1}^{(j)})$
    end
    $s_t^{(j)} \sim \mathcal{N}(\mu_t^{(j)}, \Sigma_t^{(j)}) \quad \theta_t^{(j)} = \{\mu_t^{(j)}, \Sigma_t^{(j)}\}$
    switching dynamical system [1]
    $[\hat{x}_t(\theta_{1:t}^{(j)}), \hat{\Sigma}_x(\theta_{1:t}^{(j)}), n_t^{(j)}, \Sigma_{n_t}^{(j)}]$
      $= EKF(n_{t-1}^{(j)}, \Sigma_{n_{t-1}}^{(j)}, \theta_t^{(j)}, x_t)$
    calculate sample weights
    $\tilde{\omega}_t^{(j)} \propto \omega_{t-1}^{(j)} \mathcal{N}(\hat{x}_t(\theta_{1:t}^{(j)}), \hat{\Sigma}_x(\theta_{1:t}^{(j)}))$
  end
  $\Sigma_{j=1}^J \tilde{\omega}_t^{(j)} = 1$
  Compute $N_{eff} = \left\{ \sum_{j=1}^J \left(\tilde{\omega}_t^{(j)} \right)^2 \right\}^{-1}$
  if $N_{eff} \leq \eta$, resample the particles and $\omega_t^{(j)} = \frac{1}{J}$
  $\hat{n}_t = \Sigma_{j=1}^J \omega_t^{(j)} n_t^{(j)} \quad \hat{s}_t = \Sigma_{j=1}^J \omega_t^{(j)} s_t^{(j)}$
  Detection of speech/non-speech frame
end

---

3. SIMULATIONS

3.1. Simulation Setup

We compare three processing schemes: first one is a method proposed by Fujimoto *et al.* [1] where Vector Taylor Series (VTS) method and MMSE are not employed (conventional) [2], second one is the proposed method using EKF and third one is the proposed method using UKF. [3] Two types of data set are made for evaluations. First one is clean speeches recorded in a sound proof chamber, second one is noisy speeches which are artificially generated by adding 3 types of noises. Noise data are taken from "Sound Scene Database in Real Acoustical Environment" [11]. We employ white noise, particle noise and shaver noise. Then, these noises are artificially added to clean speeches with SNRs from 0 to 9dB. 100 utterances uttered by 4 males and 2 females are used for this evaluation. The contents of the utterances are TV controlling commands, e.g. "volume up", "turn off" and so on. The total number of evaluation data for each SNR is 3,600 short phrases.

[2]These processings require large processing costs, so we estimted the clean speech as $\hat{s}_t = \Sigma_{j=1}^J \omega_t^{(j)} s_t^{(j)}$ after the particle filtering step.

[3]We compared the proposed method with the Spectral Subtraction method. We cannot obtain the improvement of the speech recognition rate.

GMM with 256 mixture distributions is trained using 500 utterances uttered by 3 males and 2 females.

An acoustic model for speech recognition is developed using the Acoustical Society of Japan (ASJ) continuous speech corpus [12]. The feature parameters for the acoustic model is composed of 39 Mel Frequency Cepstral Coefficients (MFCCs) with 13 MFCCs (with zero-th MFCC) and their first and second order derivatives. At the feature extraction step, Cepstral Mean Subtraction (CMS) is applied to each sentence.

Parameters for the particle filtering is as follows: w_t is set to $\Sigma_w = 0.1$, u_t is set to $\Sigma_u = 0.0001$ and z_t is set to $\Sigma_z = 1$. The number of particles is 100. Parameters for the Polyak averaging and feedback have four states respectively, e.g. $\alpha_p = \{0.05, 0.1, 0.15, 0.2\}$, $\beta_p = \{0.5, 1.0.1.5, 2.0\}$ and $T_p = \{5, 10, 15, 20\}$. Moreover, a parameter for the switching dynamical system is $\gamma = 0.5$ [1]. μ_N and Σ_N are calculated from the first 5 obserbed samples.

The parameter α for DPM is set to larger value than the length of utterance T. [4]

We have no *a priori* information on the speech signal distribution. The hyperparameters being not known a priori, a simple estimation process is introduced. This estimation bases on the difference between the received signal and the received signal estimated using the estimated clean signal at $t-1$ and the estimated noise signal at t. That is to say, at the time t, the clean signal is estimated roughly as follows:

$$\begin{aligned} \tilde{s}_t^{(j)} &= s_{t-1}^{(j)} + \Delta s^{(j)} + z_t^{(j)} \\ \Delta s_t^{(j)} &= x_t - (\bar{s}_{t-1}^{(j)} + \log(1 + \exp(\bar{n}_t^{(j)} - \bar{s}_{t-1}^{(j)}))) \end{aligned} \tag{4}$$

where $\bar{n}_t^{(j)}$ is obtained from the Polyak averaging [1], $\bar{s}_t^{(j)}$ is obtained from the average over the 5 past frames and $z_{t-1}^{(j)} \sim \mathcal{N}(0, \Sigma_z)$. $\Delta s^{(j)}$ is determined from the past errors and the effect of the past error decays according to the exponential function. Then, the mean vector and covariance matrix of $\tilde{s}_t^{(j)}$ over all particles are calculated and we regard these values as μ_0 and Λ_0 of hyperparameters. Then $\kappa_0 = 1$ and $\nu_0 = 500$ are used for this simulation.

3.2. Results

Firstly, the noise and clean speech estimation results are shown. Figure 1 shows one example of the noise and speech tracking results by the proposed method using UKF. The abscissa is the number of frame and the ordinate is the average energy of filter bank output in the log spectral domain. The proposed method can track the noise sequence in case SNR is 9dB.

Secondly, the speech recognition rates are compared. Evaluations are performed using speech recognition decoder "Julian" [13]. Clean speeches are recorded in a sound proof chamber using a close contact microphone. Table 1 shows

[4]We have decided the parameter α from the preliminary evaluation.

Table 1. Speech Recognition Rate for Noisy Data (%)

	white				shaver				particle			
	no processing	EKF	UKF	GMM	no processing	EKF	UKF	GMM	no processing	EKF	UKF	GMM
0dB	3.0	20.5	16.2	3.0	8.3	31.3	29.7	7.0	7.8	27.0	23.0	10.2
3dB	16.5	54.0	49.5	11.2	37.5	55.2	56.7	17.0	30.8	50.2	48.0	21.8
6dB	51.7	76.8	76.5	33.0	64.8	69.8	74.5	37.7	62.8	67.7	68.3	39.0
9dB	80.2	87.3	87.8	52.8	81.5	78.7	83.0	52.2	85.0	81.7	82.5	53.5

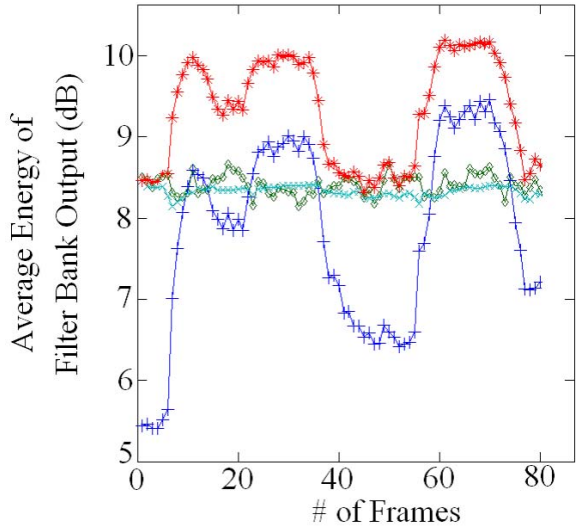

Fig. 1. Tracking result of the proposed method using UKF (in case where a noise signal is a particle noise and SNR is 9dB) $*$: received (observed) signal, $\diamondsuit$: true noise signal, $\times$: estimated noise signal, $+$: estimated clean signal

speech recognition rates. In this table, the speech recognition rate for three types of noise data (white noise, shaver noise, particle noise) are shown. Moreover, for each noise data, there are the speech recognition rates of four processing schemes (no processing, proposed method using EKF, UKF and conventional method using GMM). From this table, it can be found that speech recognition rates are improved using the proposed method using EKF in case the SNRs are 0, 3 and 6dB. Applying UKF, we can get better results than those of EKF in case of high SNR.

The speech recognition rate by the conventional method is lower than even that with no processing. The reason is that the time allocated to the GMM learning is not enough long. [5]

4. CONCLUSION

In this paper, we proposed a method for modeling the clean speech distribution using DPM and noise sequence using particle filtering. Our proposed method realizes better noise estimation accuracy than the method using inaccurate GMM. In the evaluation using speech recognition, our proposed method can improve the speech recognition rate in the SNRs 0dB, 3dB, 6dB and 9dB except for the particle noise.

[5] Although this is one reason, we obtained better speech recognition rate by employing the conventional method with VTS and MMSE on the limited data set. The required processing time became 10 times more than that of the convential method without VTS and MMSE.

5. REFERENCES

[1] M. Fujimoto *et al.*, "Sequential non-stationary noise tracking using particle filtering with switching dynamical system," Proc. ICASSP2006, pp. 769-772, May 2006.

[2] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to nonparametric problems.," Annals of Statistics, 2, pp.1152–1174, 1974.

[3] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems.," Annals of Statistics 1, pp.209–230, 1973.

[4] M. D. Escobar *et al.*, "Bayesian Density Estimation and Inference using Mixtures," Journal of the American Statistical Association, Vol. 90, No. 430, 1995.

[5] A. Kottas, "Dirichlet Process Mixtures of Beta Distributions, with Applications to Density and Intensity Estimation.," Proc. of the Workshop on Learning with Nonparametric Bayesian Methods, 23rd ICML, 2006.

[6] F. Caron *et al.*, "Bayesian Inference for Dynamic Models with Dirichlet Process Mixtures.", Proc. of FUSION'06, Florence, Italia, July 10-13, 2006.

[7] J. C. Segura *et al.*,"Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using AURORA II Database and Tasks," Proc. EuroSpeech '01, Vol. I, pp. 221-224, Sept. 2001.

[8] D. Blackwell *et al.*, "Ferguson distributions via polya urn schemes.," Annals of Statistics, 1:353–355, 1973.

[9] A. Gelman *et al.*, "Bayesian data analysis", Chapman and Hall, 1995.

[10] K. Ota *et al.*, "Bayesian Inference for Speech Density Estimation by the Dirichlet Process Mixture", Journal of SIC., vol. 16, No. 3, pp. 227–244, 2007.

[11] http://tosa.mri.co.jp/sounddb/indexe.htm

[12] http://www.mibel.cs.tsukuba.ac.jp/jnas/instruct.html

[13] http://julius.sourceforge.jp/