Online Estimation of Speech and Noise Vectors

Desh Raj

December 2019

1 Abstract

In this note, I describe a method for estimating the utterance-level means of the speech frames and noise frames when only a part of the utterance has been observed.

2 Preliminaries

We assume the following: every utterance i is represented as a matrix $d \times N_i$, where N_i is the number of frames in the utterance and d is the length of vector for each frame (usually $d = 40$ for high-resolution MFCC features). At training time, we observe all N_i frames for the utterance, but at test time, since we perform the computation in an online fashion, we only observe $T_i < N_i$ frames at any time t.

Let $\boldsymbol{x}_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_{T_i}})$ denote the sequence of observed frames and $\boldsymbol{s}_i = (s_{i_1}, s_{i_2}, \ldots, s_{i_{T_i}})$ denote the decision of a speech activity detection system which classifies each frame as speech or noise, i.e., $s_j = 1$ if frame is a speech frame, and 0 otherwise.

Let us assume that the speech and noise frames for utterance i are generated from multivariate normal distributions $\mathcal{N}(\boldsymbol{\mu}_{s_i}, \boldsymbol{\Sigma}_{s_i})$ and $\mathcal{N}(\boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{n_i})$. If we have seen the whole utterance, it is easy to estimate the parameters using maximum likelihood estimation. However, our task is "online," which means that we do not know the whole utterance.

Furthermore, the means of speech and noise frames $\boldsymbol{\mu}_{s_i}$ and $\boldsymbol{\mu}_{n_i}$ are not independent, since speech frames also contain background noise.

3 Method

Our task is to compute

$$\hat{\boldsymbol{\mu}}_{s_i}, \hat{\boldsymbol{\mu}}_{n_i} = \arg\max p(\boldsymbol{\mu}_{s_i}, \boldsymbol{\mu}_{n_i} | \boldsymbol{x}_i, \boldsymbol{s}_i, \pi), \tag{1}$$

where π is the prior. We start by expanding the joint likelihood as follows.

$$p(\boldsymbol{x}_i, \boldsymbol{\mu}_{s_i}, \boldsymbol{\Sigma}_{s_i}, \boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{n_i} | \boldsymbol{s}_i, \pi) = \prod_{t=1}^{T_i} p(x_{i_t} | s_{i_t}, \boldsymbol{\mu}_{s_i}, \boldsymbol{\Sigma}_{s_i}, \boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{n_i}) p(\boldsymbol{\mu}_{s_i} | \boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{s_i}) p(\boldsymbol{\Sigma}_{s_i} | \pi) p(\boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{n_i} | \pi), \tag{2}$$

where

$$p(x_{i_t} | s_{i_t}, \boldsymbol{\mu}_{s_i}, \boldsymbol{\Sigma}_{s_i}, \boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{n_i}) = \begin{cases} \mathcal{N}(x_{i_t} | \boldsymbol{\mu}_{s_i}, \boldsymbol{\Sigma}_{s_i}), & \text{if } s_i = 1, \\ \mathcal{N}(x_{i_t} | \boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{n_i}), & \text{if } s_i = 0 \end{cases} \tag{3}$$

$$p(\boldsymbol{\mu}_{s_i} | \boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{s_i}) = \mathcal{N}(x_{i_t} | a + B\boldsymbol{\mu}_{n_i}, b\boldsymbol{\Sigma}_{s_i}) \tag{4}$$

$$p(\boldsymbol{\Sigma}_{s_i} | \pi) = \mathcal{IW}(\boldsymbol{\Sigma}_{s_i} | \delta_s, \Phi_s) \tag{5}$$

$$p(\boldsymbol{\mu}_{n_i}, \boldsymbol{\Sigma}_{n_i} | \pi) = \mathcal{N}(\boldsymbol{\mu}_{n_i} | \boldsymbol{\mu}_n, c\boldsymbol{\Sigma}_{n_i}) \mathcal{IW}(\boldsymbol{\Sigma}_{n_i} | \delta_n, \Phi_n), \tag{6}$$

where $\mathcal{IW}$ is the Inverse-Wishart prior.

3.1 Estimation of prior parameters

Since the entire training data $\mathcal{D}$ is known, we can estimate the hyperparameters of the priors (equations 4, 5, and 6) from the training data using empirical Bayes estimates.

We will first estimate the parameters of the Inverse-Wishart priors for the covariance matrices. Following a similar approach as in [1], it can be shown that the empirical Bayes estimates of the parameters δ and Φ are

$$\delta = \delta^* + n, \text{and} \tag{7}$$

$$\Phi = \frac{\delta^* + n}{n} S_y, \tag{8}$$

where $n = |\mathcal{D}|$, $S_y = \sum_{i=1}^{n} y_i y_i^T$, and δ^* is obtained from solving

$$\frac{d}{2} \log \frac{\delta + n}{2} - \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\delta + n - i - 1 - j} = 0. \tag{9}$$

Note that there is no closed form solution for δ^*.

From these, we can estimate δ_s, Φ_s, δ_n, and Φ_n. Then, for the Normal part in equation 6, we can write

$$\boldsymbol{\mu_n} = \sum_{i=1}^{|\mathcal{D}|} \boldsymbol{\mu_{n_i}}, \tag{10}$$

$$c\boldsymbol{\Sigma_n} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} (\boldsymbol{\mu_{n_i}} - \boldsymbol{\mu_n})(\boldsymbol{\mu_{n_i}} - \boldsymbol{\mu_n})^T, \tag{11}$$

where $\boldsymbol{\Sigma_n} = \delta_n \Phi_n$ is the mean of the $\mathcal{IW}$ prior. Similarly, for equation 4, we have

$$a + B\boldsymbol{\mu_n} = \sum_{i=1}^{|\mathcal{D}|} \boldsymbol{\mu_{s_i}}, \tag{12}$$

$$c\boldsymbol{\Sigma_s} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} (\boldsymbol{\mu_{s_i}} - \boldsymbol{\mu_s})(\boldsymbol{\mu_{s_i}} - \boldsymbol{\mu_s})^T, \tag{13}$$

where $\boldsymbol{\Sigma_s} = \delta_s \Phi_s$.

3.2 Computation of online means $\hat{\boldsymbol{\mu}}_{s_i}$ and $\hat{\boldsymbol{\mu}}_{n_i}$

At test time, we have to solve for equation 1. For sake of simplicity, we assume that the covariances $\boldsymbol{\Sigma_{s_i}}$ and $\boldsymbol{\Sigma_{n_i}}$ are fixed ($\boldsymbol{\Sigma_s}$ and $\boldsymbol{\Sigma_n}$, respectively), and does not need to be estimated.

References

[1] LR Haff. Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, pages 586–597, 1980.