Lab 05 probabilistic noise suppression methods

개요: fixed threshold로 잡음구간 (non-speech period)를 추정하면

① 입력 gain에 영향받는다.

② noise의 상대적인 크기 (SNR)의 변화에 대처하지 못한다

③ 음성구간으로 잘못 추정하면 음성도 차감될 수 있다

입력 신호에 따라 유연하게 대처하기 위해서 (e.g. adaptive thresholding)

확률 모델을 사용하는 방법을 연습한다.

* 이번 Lab부터는 VAD (voice activity detection)을 주로 수행한다.

- EPD (end point detection) : 음성구간의 시작과 끝을 추정

- VAD : 음성의 유무 (active/deactive)를 주로 frame별로 (주로 10ms 단위)
결정한다. 0/1 의 hard decision, $P(voice / y(t))$ 의 soft decision이
가능하며 확률 모델을 사용한 경우 hard decision은
soft decision을 thresholding 하며 얻을 수 있다

- $P(voice | y(t))$: posterior voicing probability of $y(t)$ being observed
$y(t)$ 가 관측되었을 때 후험 확률

- EPD는 VAD 결과를 후처리 (post processing)하며 얻을 수 있다.
thresholding, median filtering, 구간 길이 검증

I. noise spectrum estimation and suppression by probabilistic VAD

a. 적절한 확률모델로 voicing probability $P(V|y)$ 구함.

b. 각 frame별로 noise 확률 계산, $1 - P(V|y_k)$, $y_k = [y_{k1} \dots y_{kN}]$ frame k

c. (hard decision)

$$1 - P(V|y_k) \geq \theta_{noise} \rightarrow I(k) = 1$$
$$< \theta_{noise} \rightarrow I(k) = 0 \Big) \quad k = 1 \dots K \text{ #frames}$$

$$|\hat{N}(\omega)|^2 = E[|N(\omega)|^2] = \frac{\sum_{k=1}^{K} I(k) |Y(k,\omega)|^2}{\sum_{k=1}^{K} I(k)}$$

d. (soft decision) MAP estimation (maximum a posteriori estimation)
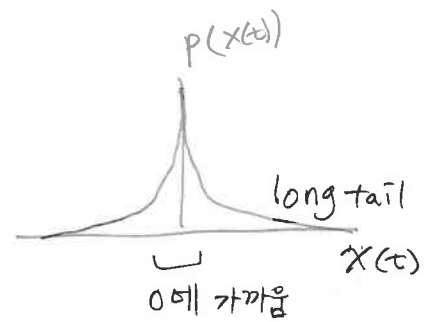noise spectrum을 posterior probability로 weighted summation

$$|\hat{N}(\omega)|^2 = \frac{\sum_{k=1}^{K} P(noise|y_k) |Y(k,\omega)|^2}{\sum_{k=1}^{K} P(noise|y_k)} \quad , \quad \begin{array}{l} P(noise | y_k) \\ = 1 - P(voice | y_k) \end{array}$$

e. lab04의 time-domain FIR Wiener filtering 이용

II. time-domain VAD

a. 음성신호의 분포는 0이 많은 sparse distribution
Laplacian distribution을 따른다고 알려져 있다.
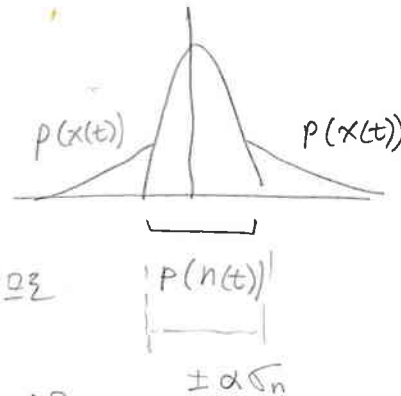(exponential decay)

거의 0

$P(x(t))$

long tail

$x(t)$

0에 가까움

$$P(x(t)) \sim \frac{1}{\sqrt{2}\sigma_x} \exp\left(-\sqrt{2}\frac{|x-u_x|}{\sigma_x}\right)$$

* 가정: noise의 크기가 speech보다 작다.

b. $y(t) = x(t) + n(t)$
noise의 분포는 Gaussian으로 가정

$$P_n(n(t)) \sim \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left(-\frac{(x-u_n)^2}{2\sigma_n^2}\right)$$

$P(x(t))$ $P(x(t))$

$P(n(t))$

$\pm \alpha \sigma_n$

speech의 분포도 sharp peak가 사라지므로
Gaussian으로 modeling

$$P_x(x(t)) \sim \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x-u_x)^2}{2\sigma_x^2}\right)$$

dual Gaussian mixture model with mean 0 ($u_x = u_n = 0$)

$$P(y(t)) = P_{voice} \cdot P_x(y(t)) + (1-P_{voice}) \cdot P_n(y(t)) \quad --- ①$$

voicing probability는? - posterior prob of $y(t)$

$$P(voice \mid y(t)) = \frac{P(voice, y(t))}{P(y(t))} = \frac{P(v) \cdot P(y(t)|v)}{P(y(t))} = \frac{P_{voice} \cdot P_x(y)}{P(y)} \quad ②$$

$$P(noise \mid y(t)) = \frac{P_{noise} \cdot P_n(y(t))}{P(y(t))} = \frac{(1-P_{voice})P_n(y)}{P(y)} \quad --- ③$$

* $\underline{P(v|y) + P(n|y) = 1} \rightarrow \underline{P(n|y) = 1 - P(v|y)}$

①, ③을 조합하면

$$P(noise \mid y(t)) = \frac{(1-P_v) \cdot P_n(y(t))}{P_v \cdot P_x(y) + (1-P_v) \cdot P_n(y)} \quad \gtrless 0.5 \begin{array}{l} \nearrow \text{noise} \\ \searrow \text{speech} \end{array}$$

soft decision   hard decision

c. sample 별로 하지 않고 frame별로 decision
let $u[k] \equiv \sqrt{\sum_{t \in frame\ k} y(t)^2}$ (frame energy. 평균을 위해 N+로 나누어도 된다. 어차피 상수)

$P(frame\ k\ is\ noise) = P(noise \mid u[k])$, 당연히 학습도 frame energy로

Ⅱ. time-domain VAD (continued)

d. dual Gaussian mixture training

EM (expectation maximization) algorithm 사용

<offline training>

→ √를 쓰우는 것은 나중에 $u[k]^2$으로 학습계산 하기 때문 (연산줄이)

① compute $u[k] = \sqrt{\sum_{t \in \text{frame } k} y(t)^2}$ for all $k = 1 \dots K$

② Sort, and split ⎡ lower $1 \sim \frac{K}{2}$ frames → compute σ_n^2
                   ⎣ upper $\frac{K}{2} \sim K$ frames → σ_x^2

③ $u_n = u_x = 0$, $P_{voice} = 0.5$

④ compute $p_x(u[k]) = \dfrac{1}{\sqrt{2\pi}\,\sigma_x} \exp\left(-\dfrac{u[k]^2}{2\sigma_x^2}\right)$

$p_n(u[k]) = \dfrac{1}{\sqrt{2\pi}\,\sigma_n} \exp\left(-\dfrac{u[k]^2}{2\sigma_n^2}\right)$

$p(\text{voice} \mid u[k]) = \dfrac{P_v\, p_x(u[k])}{P_v\, p_x(u[k]) + (1-P_v)\, p_n(u[k])}$

⑤ update $P_v \Leftarrow \dfrac{\sum_{k=1}^{K} P(v \mid u[k])}{K}$

$\sigma_x^2 = \dfrac{\sum_{k=1}^{K} P(v \mid u[k]) \cdot u[k]^2}{\sum_{k=1}^{K} P(v \mid u[k])}$

$\sigma_n^2 = \dfrac{\sum_k P(\text{noise} \mid u[k]) \cdot u[k]^2}{\sum_k P(\text{noise} \mid u[k])} = \dfrac{\sum_k (1 - P(v \mid u[k])) \cdot u[k]^2}{\sum_k (1 - P(v \mid u[k]))}$

e. noise spectrum estimation

$|\tilde{N}(w)|^2 = E_N[N(k,w)] = \dfrac{\sum_k P(\text{noise} \mid u[k]) \cdot u[k]^2}{\sum_k P(\text{noise} \mid u[k])} = \dfrac{\sum_k (1 - P(v \mid u[k])) \cdot u[k]}{\sum_k (1 - P(v \mid u[k]))}$ (soft)

$= \dfrac{\sum_k (1 - I_x[k]) \cdot u^2[k]}{\sum_k (1 - I_x[k])}$ (hard)

where $I_x[k] = \begin{cases} 1 & \text{if } P(v \mid u[k]) \geq 0.5 \\ 0 & \text{if } \quad '' \quad < 0.5 \end{cases}$

$I_x[k]$는 voice의 posterior 확률이 더 클때 1이 되는 indicator function

Ⅲ. frequency domain VAD using Rayleigh mixture model

만약 noise 특성이 (예를 들어) 저주파에 집중되어 있고 amplitude가 매우 크다면 time-domain VAD에 매우 불리함

(예) n x $y = x + n$

noise 크기가 매우 크지만 주파수 영역에서는 narrow band

이와 같은 colored noise에 대해서 신뢰성 있는 VAD 결과를 얻기 위해 frequency domain에서 확률 모델을 사용하여 본다.

$Y(k, w)$: frame k의 주파수 w 성분

$|Y(k,w)|^2 = Y(k,w) \, Y^*(k,w)$ — $\circ^*$는 complex conjugate (켤레 복소수)

$|Y(k,w)| \sim$ Rayleigh(σ), pdf : $f_x(x|\sigma) = \dfrac{x}{\sigma^2} \exp\left(-\dfrac{x^2}{2\sigma^2}\right)$

Rayleigh Distribution

두개의 Random Variable X, Y가 평균이 0이고 같은 variance σ^2의 Gaussian distribution을 따르면 $(X \sim N(0, \sigma^2), \ Y \sim N(0, \sigma^2))$ random variable $R = \sqrt{X^2 + Y^2}$ 은 Rayleigh distributed (wikipedia)

*단, $X \perp Y \Leftrightarrow E[XY] = 0$, 두 변수 X와 Y는 independent (i.i.d.) (uncorrelated)

따라서 $Y(k,w) = A + jB$ 이고 A, B가 independent Gaussian with same variance σ^2

$\Rightarrow |Y(k,w)| \sim f_x(x|\sigma) = \dfrac{x}{\sigma^2} \exp\left(-\dfrac{x^2}{2\sigma^2}\right)$

① $f_x(0|\sigma) = \dfrac{0}{\sigma^2} \cdot \exp(-0) = 0.1 = 0$

② $x < 0$는 정의되지 않음 $f_x(x|\sigma) = 0$ for $x < 0$

③ $y = x$ 와 $y = \exp(-x^2)$ 이 곱해진 형태

④ $x = \sigma$는 mode (local peak) ← 중요함 !!

⑤ parameter는 σ^2 하나밖에 없으며 (variance는 아님) 다음과 같이 추정한다.

$\hat{\sigma}^2 = \dfrac{1}{2N} \sum_{i=1}^{N} x_i^2$ maximum likelihood, unbiased estimation

⑥ $E[X] = \sigma\sqrt{\dfrac{\pi}{2}}$

$Var[X] = E[(X - E_x)^2] = \dfrac{4-\pi}{2}\sigma^2$, 즉 σ^2은 variance가 아님

III-2. frequency domain VAD part 2

multivariate extension of Rayleigh distribution

* 매우 복잡하다. 따라서 다음과 같이 변수들이 (dimension들이) independent하다고 가정.

$$X = [x_1 \ x_2 \ldots x_D]$$

$$f_x(x \mid \underset{\underset{\text{covariance matrix}}{\uparrow}}{\Sigma}) \ \widehat{=} \ f_{x_1}(x_1 \mid \sigma_1^2) \cdot f_{x_2}(x_2 \mid \sigma_2^2) \cdots f_{x_D}(x_D \mid \sigma_D^2)$$

$$= \prod_{d=1}^{D} f_{x_d}(x_d \mid \sigma_d^2) \quad \text{각 individual pdf들의 곱으로 표현}$$

$$= \prod_{d=1}^{D} \frac{x_d}{\sigma_d^2} \cdot \exp\left(-\frac{x_d^2}{2\sigma_d^2}\right) = \left(\prod_d \frac{x_d}{\sigma_d^2}\right) \cdot \exp\left(-\sum_d \frac{x_d^2}{2\sigma_d^2}\right)$$

$$\log f_x(x \mid \Sigma) = \sum_d \log x_d - \sum_d 2\log\sigma_d - \sum_d \frac{x_d^2}{2\sigma_d^2}$$

$$Y(k,w) = X(k,w) + N(k,w) \cdot \quad \begin{array}{l}\text{X와 N은 independent} \\ \text{(i.d.d.) distinctively distributed}\end{array}$$

$$|Y(k,w)| = |X(k,w) + N(k,w)| \implies \text{dual Rayleigh mixture model 적용.}$$

 * 엄밀히 말하면 Y는 두 complex Random variables X와 N의 합의 절대값
  mixture model은 X와 N이 번갈아 관측되는 것을 가정하므로
  오류가 있다. 하지만 speech는 sparse하므로 적용해 보도록 한다

* $\sigma_x^2 > \sigma_n^2$ 으로 가정한다

$$f_Y(y) \simeq P_x f_x(y \mid \sigma_x^2) + P_N f_N(y \mid \sigma_n^2)$$

$$= P_x \frac{y}{\sigma_x^2} \exp\left(-\frac{y^2}{2\sigma_x^2}\right) + P_N \frac{y}{\sigma_n^2} \exp\left(-\frac{y^2}{2\sigma_n^2}\right)$$

$$= y\left(\frac{P_x}{\sigma_x^2} \exp\left(-\frac{y^2}{2\sigma_x^2}\right) + \frac{P_N}{\sigma_n^2} \exp\left(-\frac{y^2}{2\sigma_n^2}\right)\right)$$

 * posterior probability

$$P(voice \mid y) = \frac{P_x f_x(y \mid \sigma_x^2)}{f_Y(y)} = \frac{P_x f_x(y \mid \sigma_x^2)}{P_x f_x(y) + P_N f_N(y)}$$

$$P(Noise \mid y) = \frac{P_N f_N(y)}{P_x f_x(y) + P_N f_N(y)} = 1 - P(voice \mid y)$$

 * EM algorithm으로 학습한다.

 * $\sigma_x^2 < \sigma_N^2 \rightarrow$ swap (학습하다가 일어날 수 있다).

 * 학습이 끝나면 $E[|\tilde{N}(w)|^2] = \sigma_N^2$ (noise Rayleigh의 mode)

Ⅲ-3. frequency domain VAD part3

$|\tilde{N}(w)|^2$ 을 주파수 w 마다 Rayleigh mixture model을 하지 않고 multivariate Rayleigh로 추정.

let $y(k) = \left[|Y(k,w_1)| \; |Y(k,w_2)| \; \ldots \; |Y(k,w_D)| \right]^T$

주파수 성분들의 절대값들로 구성된 column vector

speech와 noise의 pdf는 각
주파수 값들이 independent하다고 가정하고 각 dimension의 scalar pdf의 곱으로 표현

$$f_x(y) = f_{x_1}(y_1) \cdot f_{x_2}(y_2) \ldots f_{x_D}(y_D) = \prod_d f_{x_d}(y_d)$$

$$= \prod_d \frac{y_d}{\sigma_{xd}^2} \cdot \exp\left(-\frac{y_d^2}{2\sigma_{xd}^2}\right) = \left(\prod_d \frac{y_d}{\sigma_{xd}^2}\right) \cdot \exp\left(-\sum_d \frac{y_d^2}{2\sigma_{xd}^2}\right)$$

$$f_n(y) = \prod_d \frac{y_d}{\sigma_{nd}^2} \cdot \exp\left(-\frac{y_d^2}{2\sigma_{nd}^2}\right) = \left(\prod_d \frac{y_d}{\sigma_{nd}^2}\right) \cdot \exp\left(-\sum_d \frac{y_d^2}{2\sigma_{nd}^2}\right)$$

* computation of multivariate pdf using log

$$f_x(y) = \exp\left(\sum_d \log y_d - 2\sum_d \log \sigma_{xd} - \sum_d \frac{y_d^2}{2\sigma_{xd}^2}\right)$$

exponential function을 매우 많이 곱하면 수가 많이 작아져서 underflow가 날 수 있다. 따라서 곱하기를 줄이고 더하기로 대치한다.

$$f_n(y) = \exp\left(\sum_d \log y_d - 2\sum_d \log \sigma_{nd} - \sum_d \frac{y_d^2}{2\sigma_{nd}^2}\right)$$

* dual mixture multivariate Rayleigh

$$P(voice|y) = \frac{P_x \cdot f_x(y)}{P_x \cdot f_x(y) + P_n \cdot f_n(y)} \triangleq \gamma_x(y)$$

효율적인 계산을 위해서 분모/분자를 $P_x \cdot f_x(y)$로 나눔

$$= \frac{1}{1 + \frac{P_n}{P_x} \cdot \frac{f_n(y)}{f_x(y)}}$$

(error correction)
$n \Leftrightarrow x$

$$\log \frac{f_n(y)}{f_x(y)} = \sum_d \left(\log y_d - 2\log \sigma_{nd} - \frac{y_d^2}{2\sigma_{nd}^2} - \log y_d + 2\log \sigma_{xd} + \frac{y_d^2}{2\sigma_{xd}^2}\right)$$

$$= \sum_d \left(2\log \frac{\sigma_{xd}}{\sigma_{nd}} + \frac{y_d^2}{2}\left(\frac{1}{\sigma_{xd}^2} - \frac{1}{\sigma_{nd}^2}\right)\right) \triangleq \mathfrak{z}_x(y)$$

⇒ log 1번, exp 1번으로 $\gamma_x(y)$ 계산 가능

$$\gamma_x(y) = \frac{1}{1 + \frac{P_n}{P_x} \cdot \exp \mathfrak{z}_x(y)}$$

$$\gamma_n(y) = p(noise|y) = 1 - p(voice|y) = 1 - \gamma_x(y)$$

II-4. frequency domain VAD using Rayleigh mixture model   part 4

< EM algorithm >

① Initialization

~~처음 10~30개의~~ frame으로 energy 큰것 절반 → $f_x(y)$ Initial parameters

전체   작은것 절반 → $f_n(y)$ Initial parameters

② Expectation

compute $\gamma_x(y(k))$ for all $k = 1 \ldots \#frames$

③ maximization

update σ_{xd} and σ_{nd}, probabilistically (maximum likelihood estimate)

$$\sigma_{xd}^2 = \frac{\sum_{k=1}^{K} \gamma_x(y(k)) \circ y_d^2(k)}{2 \sum_{k=1}^{K} \gamma_x(y(k))} \quad , \quad \sigma_{nd}^2 = \frac{\sum_{k=1}^{K} (1-\gamma_x(y(k)) \cdot y_d^2(k)}{2 \sum_{k=1}^{K} (1-\gamma_x(y(k)))}$$

for all $d = 1 \ldots D$

④ Repeat 3~4.

< Noise Spectrum Estimation >

Nois Rayleigh의 mode 위치인 σ_{nd} 로 예측

$$\boxed{|\tilde{N}(\omega_d)|^2 = \sigma_{nd}^2}$$

* Implementation tips.

① $\sigma_n < \sigma_x$ 가 보장되도록 유지

② $\sum \gamma_x > 0$, $\sum(1-\gamma_x) > 0$ 인지 확인하고 너무 작으면 update 하지 않음
   (이미 diverged)

③ $|Y(k,\omega)|^2 = 0$ 이면 Wiener filter gain 계산할 때
   "divide by zero" exception 발생.
   $|Y(k,\omega)|^2 \Leftarrow \max(\varepsilon_y, |Y(k,\omega)|^2)$ 등의 방법 사용.

④ $y(t)$에 매우 작은 크기 ($\sigma = 10^{-10} \sim 10^{-6}$) 의 Gaussian random noise를
   더해주는 방법도 있다.

< Online EM algorithm >

실시간으로 들어오는 입력에는 offline training 불가능, 수정 필요

① Initialization : 초기 10~30 frame으로 초기화, Wiener filtering 하지 않음

② Expectation at current frame

compute $\gamma_x(y(k))$ for current frame k.

③ Adaptive maximization

$$\sigma_{xd}^2 = (1-\alpha \gamma_x(k)) \sigma_{xd}^{2 \,(old)} + \frac{1}{2} \cdot \alpha \gamma_x(k) \cdot y_d^2(k) . \quad 0 \leq d \leq 1$$

$$\sigma_{nd}^2 = (1-\alpha(1-\gamma_x(k))) \sigma_{nd}^{2 \,(old)} + \frac{1}{2}\alpha(1-\gamma_x(k)) \cdot y_d^2(k)$$
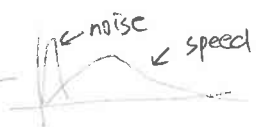
④ repeat ②-③

Ⅳ. frequency domain VAD using LogNormal distribution

Rayleigh 분포는 parameter가 하나밖에 없기 때문에 표현력이 떨어짐.
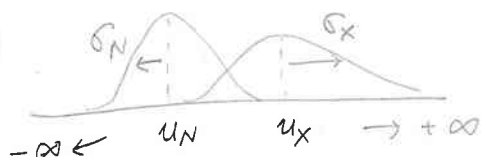특히 mixture modeling을 하면 이론과 달리 잘 안 맞을 수 있다.
또한, noise와 speech의 scale 차이가 크면 한 쪽으로 수렴될 수 있다

noise의 분포가 sharp 하지만
scale이 speech에 비해 너무 작다

< LogNormal >

Let $y(k, w) = \log |Y(k, w)|^2$, 그리고 Gaussian mixture model 사용

⇒ scale mismatch 와
표현력 $(\sigma \to u, \sigma)$ 증가

PDF

$$f_x(y) = \frac{1}{(2\pi)^{\frac{D}{2}} \cdot \sqrt{|\det \Sigma|}} \cdot \exp\left(-\frac{1}{2}(y - u_x)^T \Sigma^{-1}(y - u_x)\right)$$

u_x : mean vector, Σ : covariance matrix
너무 복잡하기 때문에 independent 가정

$$\simeq \prod_d \frac{1}{\sqrt{2\pi}\,\sigma_{xd}} \cdot \exp\left(-\frac{(y_d - u_{xd})^2}{2\sigma_{xd}^2}\right)$$

$f_n(y)$도 같은 방법으로 modeling 하고

$$\gamma_x(y) = \frac{P_x f_x(y)}{P_x f_x(y) + P_N f_N(y)} = \frac{1}{1 + \frac{P_N}{P_x}\frac{f_N(y)}{f_N(y)}}$$

EM update

$$u_x = \frac{\sum_k \gamma_x(y) \cdot y}{\sum_k \gamma_x(y)}, \quad \sigma_{xd}^2 = \frac{\sum_k \gamma_x(y) \cdot (y_d - u_{xd})^2}{\sum_k \gamma_x(y)}$$

$$u_N = \frac{\sum_k (1 - \gamma_x(y)) \cdot y}{\sum_k \gamma_k(y)} \qquad \sigma_{Nd}^2 = \frac{\sum (1 - \gamma_x(y)) \cdot (y_d - u_{Nd})^2}{\sum (1 - \gamma_x(y))}$$

$$\underline{|\hat{N}(w_d)|^2 \cong \exp(u_{Nd})} \quad \because \text{Gaussian 의 mode는 mean vector}$$