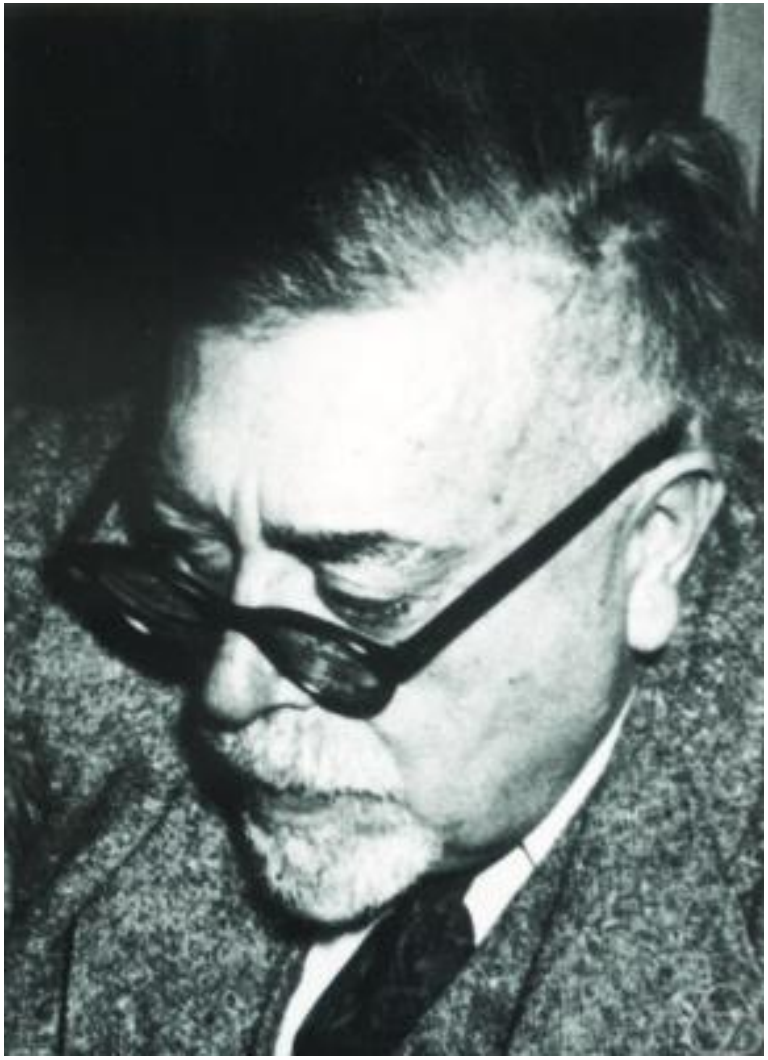
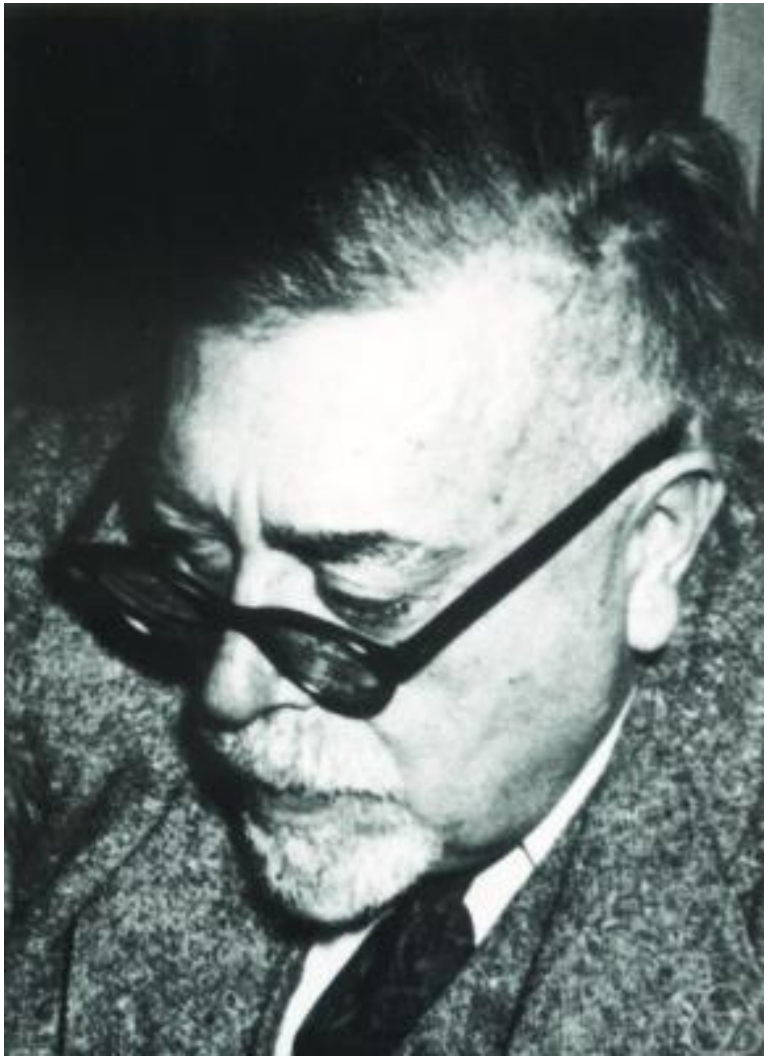
Norbert



Norbert



Norbert Wiener
1894–1964

The Wiener filter

Note

If you have the distribution of a variable, there are various ways of finding a point estimate:

- Maximum Likelihood.
- Maximum a-Posteriori.
- Expectation.

Classical (and much modern) signal processing tends to focus on MMSE (Minimum Mean Squared Error). This in turn implies expectations.
Practically, expectations involve integration and maxima involve differentiation.

Sum of two RVs

If we have two Gaussian RVs, s and n, their joint PDF is

$$p(s,n) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{s^2}{2{\sigma_s}^2}\right) \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{n^2}{2{\sigma_n}^2}\right).$$

However, we do not observe s and n together. We actually observe $t = s + n$.

So far, this is the same as the sum-of-two-Gaussians derivation.

Write as one variable and sum

Making the change of variable,

$$t = s + n, \qquad s' = s,$$

the Jacobian determinant is

$$J(s', t) = \begin{vmatrix} \dfrac{\partial s}{\partial s'} & \dfrac{\partial n}{\partial s'} \\ \dfrac{\partial s}{\partial t} & \dfrac{\partial n}{\partial t} \end{vmatrix} = 1,$$

giving

$$p(s, t) = \frac{1}{2\pi\sigma_s\sigma_n} \exp\left(-\frac{s^2}{2{\sigma_s}^2} - \frac{(t-s)^2}{2{\sigma_n}^2}\right).$$

Still the same as the sum-of-two-Gaussians.

Estimator

Now it's different:

To find an estimator, $\hat{s}$, for s, notice that

$$p(s \mid t) = \frac{p(s,t)}{p(t)}$$

calculate

$$\frac{\partial}{\partial s} \log p(s,t) = -\frac{s}{{\sigma_s}^2} + \frac{t-s}{{\sigma_n}^2} = 0$$

$$\implies \hat{s} = \frac{{\sigma_s}^2}{{\sigma_s}^2 + {\sigma_n}^2} t.$$

Same thing in the complex case

As the real and imaginary components of the complex Gaussian are independent, the Wiener filter extends trivially to the complex case. The joint PDF is

$$p(\mathfrak{s}, \mathfrak{n} \mid \sigma, \nu) = \frac{1}{\pi\sigma} \exp\left(-\frac{|\mathfrak{s}|^2}{\sigma}\right) \frac{1}{\pi\nu} \exp\left(-\frac{|\mathfrak{n}|^2}{\nu}\right),$$

and after a change of variable, this becomes

$$p(\mathfrak{s}, \mathfrak{t} \mid \sigma, \nu) = \frac{1}{\pi^2\sigma\nu} \exp\left(-\frac{|\mathfrak{s}|^2}{\sigma} - \frac{|\mathfrak{t}-\mathfrak{s}|^2}{\nu}\right).$$

Same answer

This can be solved by differentiating w.r.t. the real and imaginary parts of $\mathfrak{s}$ separately, then combining to give

$$\hat{\mathfrak{s}} = \frac{\sigma}{\sigma + \nu}\mathfrak{t}.$$

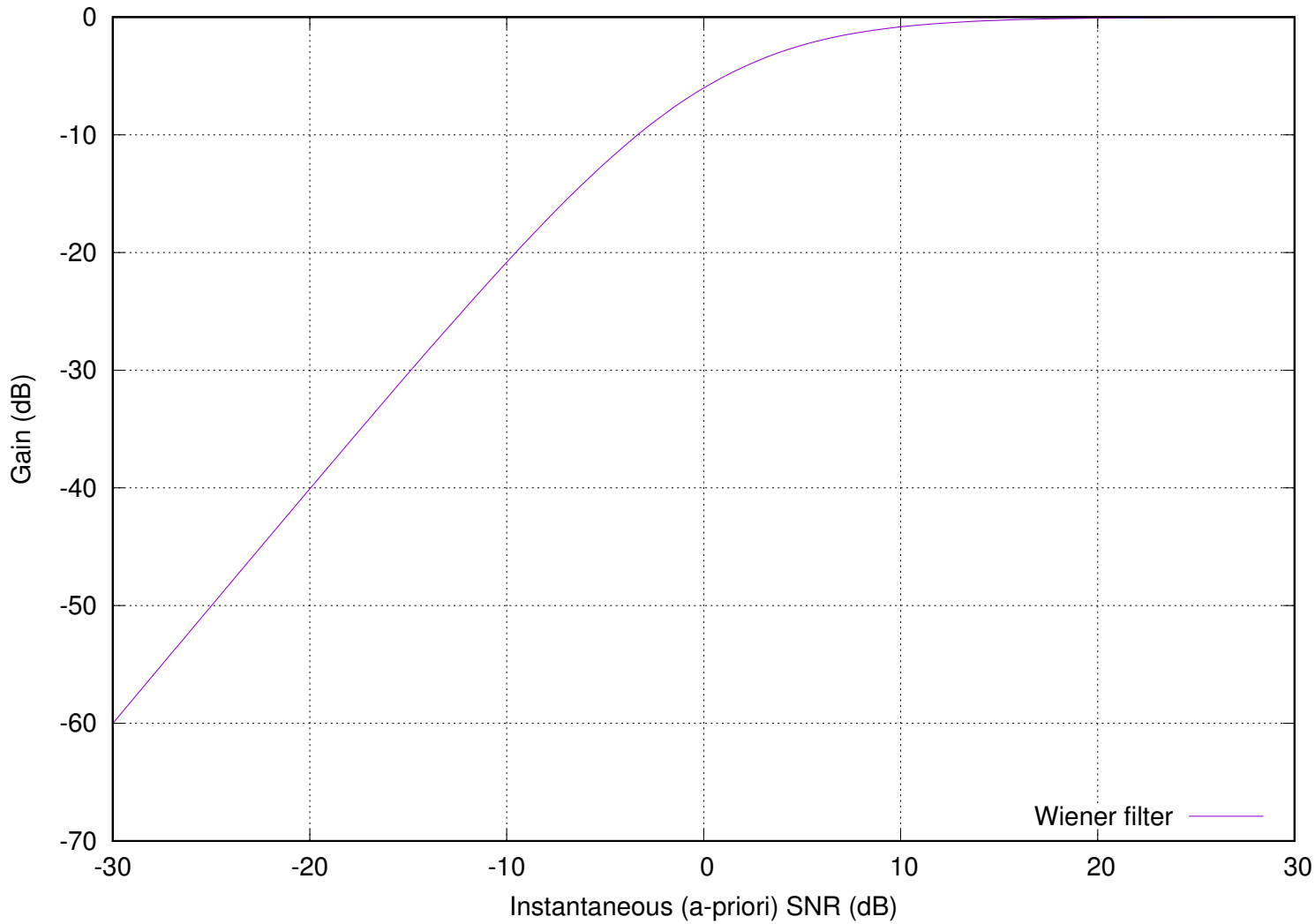
The quantity in the fraction is known as the Wiener filter.

- It's typically derived as a MMSE estimator.
- It is a suppression filter - something that scales $\mathfrak{t}$.

Notice that

$$\hat{\mathfrak{s}} = \frac{\sigma/\nu}{\sigma/\nu + 1}\mathfrak{t}.$$

Wiener as a suppression filter



Variance estimation

Estimators related to the Wiener filter require some values for the noise and speech variances.
There are two distinct cases

Noise We can assume that we know the noise variance. In practice, it can be estimated from bits of the signal where only noise is present. Was covered at the end of the last lecture

Speech The speech variance is really unknown. If we knew it, noise removal would be trivial.

Estimating (or otherwise dealing with) speech variance is non-trivial.

ML estimator of speech variance

Recall the sum-of-two-Gaussians expression:

$$p\left(t \mid \sigma, \nu\right) = \frac{1}{\pi(\sigma + \nu)} \exp\left(-\frac{|t|^2}{\sigma + \nu}\right).$$

Consider using this as a basis for estimation of the speech variance, σ. Bayes' theorem gives

$$p\left(\sigma \mid t, \hat{\nu}\right) \propto p\left(t \mid \sigma, \hat{\nu}\right) p\left(\sigma\right).$$

If we assume a flat prior $p\left(\sigma\right) \propto 1$, the problem then reduces to the maximisation

$$\hat{\sigma} = \max_{\sigma} p\left(t \mid \sigma, \hat{\nu}\right),$$

Spectral subtraction

To find the estimate, differentiate and equate to zero:

$$\hat{\sigma} = \begin{cases} |t|^2 - \hat{\nu}, & \text{if } |t|^2 > \hat{\nu}, \\ 0 & \text{otherwise.} \end{cases}$$

This is the well known Maximum Likelihood estimate.
This is an important result! It's known as "Power Spectral Subtraction", i.e., you subtract the noise in the power domain. Some authors advocate doing this in the spectral domain, but it's not clear that such an approach is model based.

Summary

- The Wiener filter is a MAP, or MMSE, estimator of a Gaussian signal.
- It requires knowledge of the underlying speech and noise variances.
- Noise can be estimated when people are not speaking.
- The ML estimator of the speech variance is Power Spectral Subtraction.

Beware: The combination of spectral subtraction and Wiener filter cancels out. You need smoothing somewhere.

The Ephraim-Malah enhancement technique

Ephraim-Malah

In 1984 and 1985, Yariv Ephraim and David Malah wrote a couple of papers on speech enhancement.
Much speech enhancement literature since then is based, to an extent, on this work.
There are two (or three) distinct contributions:

- A MMSE spectral amplitude estimator.
- The decision-directed estimator (for the speech variance).

Then later,

- A MMSE log spectral amplitude estimator.

Beware, this stuff is not Bayesian! The underlying ideas are important though.

Human hearing

The basic premise of the Ephraim-Malah approach is that human hearing is not sensitive to complex frequency, so the Wiener filter is not the right thing to use for enhancement.
It isn't clear what is the right domain for hearing

- Hynek would advocate cube root of power.

Certainly spectral amplitude or log-spectral amplitude is closer to human hearing than power.

Steve

Steve

Stephen O. Rice
1907–1986

Distribution of magnitude 1

If we begin with

$$\mathfrak{p}\left(\mathfrak{s},\mathfrak{t}\mid\sigma,\nu\right)=\frac{1}{\pi^2\sigma\nu}\exp\left(-\frac{|\mathfrak{s}|^2}{\sigma}-\frac{|\mathfrak{t}-\mathfrak{s}|^2}{\nu}\right)$$

and change variables such that both complex variables are in polar form, we find

$$\mathfrak{p}\left(|\mathfrak{s}|,|\mathfrak{t}|,\theta,\phi\mid\sigma,\nu\right)=$$
$$\frac{|\mathfrak{s}||\mathfrak{t}|}{\pi^2\sigma\nu}\exp\left(-\frac{|\mathfrak{s}|^2}{\sigma}-\frac{|\mathfrak{s}|^2+|\mathfrak{t}|^2-2|\mathfrak{s}||\mathfrak{t}|\cos(\phi-\theta)}{\nu}\right).$$

Distribution of magnitude 2

One angle can be eliminated by making the change of variable

$$\theta' = \theta,$$
$$\psi = \phi - \theta,$$

and integrating over θ', which does not appear in the expression, so just introduces a multiplier of 2π,

$$p\left(|\mathfrak{s}|, |\mathfrak{t}|, \psi \mid \sigma, \nu\right) = \frac{2|\mathfrak{s}||\mathfrak{t}|}{\pi\sigma\nu} \exp\left(-\frac{|\mathfrak{s}|^2}{\sigma} - \frac{|\mathfrak{s}|^2 + |\mathfrak{t}|^2 - 2|\mathfrak{s}||\mathfrak{t}|\cos(\psi)}{\nu}\right).$$

Distribution of magnitude 3

Integration over the angle ψ can be done by noting that the modified Bessel function

$$\mathrm{I}_0(z) = \frac{1}{\pi}\int_0^{\pi} \mathrm{d}\theta \, \exp(z\cos\theta).$$

Hence,

$$\mathrm{p}\left(|\mathfrak{s}|, |\mathfrak{t}| \mid \sigma, \nu\right) = \frac{4|\mathfrak{s}||\mathfrak{t}|}{\sigma\nu} \exp\left(-\frac{|\mathfrak{s}|^2}{\sigma} - \frac{|\mathfrak{s}|^2 + |\mathfrak{t}|^2}{\nu}\right) \mathrm{I}_0\left(\frac{2|\mathfrak{s}||\mathfrak{t}|}{\nu}\right).$$

Distribution of magnitude 4

Given that

$$\mathrm{p}\left(|\mathfrak{t}| \mid \sigma, \nu\right) = \frac{2|\mathfrak{t}|}{\sigma+\nu} \exp\left(-\frac{|\mathfrak{t}|^2}{\sigma+\nu}\right),$$

it can be shown that

$$\begin{aligned}\mathrm{p}\left(|\mathfrak{s}| \mid |\mathfrak{t}|, \sigma, \nu\right) &= \frac{\mathrm{p}\left(|\mathfrak{s}|, |\mathfrak{t}| \mid \sigma, \nu\right)}{\mathrm{p}\left(|\mathfrak{t}| \mid \sigma, \nu\right)} \\ &= \frac{2|\mathfrak{s}|}{\upsilon} \exp\left(-\frac{|\mathfrak{s}|^2 + z^2}{\upsilon}\right) \mathrm{I}_0\left(\frac{2|\mathfrak{s}|z}{\upsilon}\right),\end{aligned}$$

which is a Rician distribution with

$$\frac{1}{\upsilon} = \frac{1}{\sigma} + \frac{1}{\nu}, \qquad z = \frac{\sigma}{\sigma+\nu}|\mathfrak{t}|.$$

Rician moments

The Ephraim-Malah estimator is a MMSE estimator.
It is the first moment of the Rician.
Check Wikipedia:
`http://en.wikipedia.org/wiki/Rice_distribution`

$$\mathbb{E}(x) = \sigma\sqrt{\frac{\pi}{2}}\, L_{1/2}\left(-\frac{\nu^2}{2\sigma^2}\right)$$

where $L_n(x)$ is a Laguerre polynomial

$$L_{1/2}(x) = \exp\left(\frac{x}{2}\right)\left[(1-x)I_0\left(-\frac{x}{2}\right) - xI_1\left(-\frac{x}{2}\right)\right],$$

and $I_n(x)$ is the modified Bessel function of the first kind.
Bleargh!

Spectral magnitude estimator

It expands to:

$$|\widehat{\mathfrak{s}}| = \Gamma(1.5)\sqrt{\upsilon}\exp\left(-\frac{z^2}{2\upsilon}\right) \times \left[\left(1+\frac{z^2}{\upsilon}\right)\mathrm{I}_0\left(\frac{z^2}{2\upsilon}\right)+\frac{z^2}{\upsilon}\mathrm{I}_1\left(\frac{z^2}{2\upsilon}\right)\right]. \quad (1)$$

In my notation anyway.
The original paper uses slightly different notation. It also formulates it as a suppression filter.
i.e., something multiplied by $|\mathfrak{t}|$.

Aside

Ephraim and Malah use the first moment, $\mathbb{E}(x)$, of the Rician. Odd moments of the Rician are complicated, but even moments are much simpler, e.g.,

$$\mathbb{E}\left(x^2\right) = 2\sigma^2 + \nu^2.$$

If you estimate spectral power (expectation of squared magnitude), the expression is easier. No special functions. This is typical of (Gaussian) spectra in general; the power is the natural quantity to deal with.

Two estimates of speech variance

Say you want to estimate the underlying speech variance, consider two methods:

- Spectral Subtraction gives the ML estimate

$$\hat{\sigma} = \max(|t|^2 - \nu, 0).$$

- You can infer something from the MMSE estimator

$$\hat{\sigma} = \mathbb{E}\left(|\widehat{s}|^2\right).$$

A Bayesian might be able to combine these as prior and likelihood terms.

Decision Directed estimator

A pragmatic approach is combine them:

$$\hat{\sigma}_{n,k} = \alpha|\widehat{s}|^2_{n-1,k} + (1-\alpha)\max(|t|^2_{n,k} - \nu_{n,k}, 0).$$

i.e.,

- A linear combination of the ML and MMSE estimators.
- The MMSE is from the previous frame, ML from the current.

This is known as the "Decision Directed" estimator[1].
In practice, $\alpha \approx 0.98$, i.e., heavily biassed towards the previous frame.

[1] I don't see why.

SNR

Ephraim and Malah follow MacAulay and Malpass by working with SNR instead of variance.
Define two Signal to Noise ratios:

- The *a-priori* SNR
$$\xi = \frac{\sigma}{\nu}.$$
- The *a-posteriori* SNR
$$\gamma = \frac{|t|^2}{\nu}.$$

The a-priori SNR is a very useful quantity. The other one is less useful.

The DD estimator is then an estimator of *a-priori* SNR:

$$\hat{\xi}_{n,k} = \alpha \frac{\widehat{|S|}^2_{n-1,k}}{\nu_{n-1,k}} + (1-\alpha)\max(\gamma_{n,k} - 1, 0).$$

Notice it's not quite the same as above if the noise estimate changes with time.

Homework

Try to draw the Ephraim-Malah rule as a suppression filter.
How does it compare (graphically) to the Wiener suppression rule?
This will help you with the next lab!

Summary

- The Ephraim Malah estimator is a MMSE estimator of spectral magnitude in noise.
- It is the first derivative of a Rician distribution.
- They use the Decision-Directed estimator to estimate speech variance.
- It is likely that the DD estimator has more effect than the suppression rule.

Voice Activity Detection

Introduction

VAD: Detection of whether or not someone is speaking.
VAD is important

In ASR because it distinguishes the non-speech portions at the beginning and end of an utterance from the utterance itself. In doing this, the VAD ensures that the decoder, which is computationally intensive, only runs when necessary.

In cellphones because it saves bandwidth when someone is not talking.

These points are particularly important in embedded applications, where processing power is limited.

Hypothesis test

Define a boolean variable or hypothesis $\mathcal{H}$, which can take values 0 and 1. $\mathcal{H} = 0$ indicates non-speech and $\mathcal{H} = 1$ indicates the presence of speech. A VAD produces an estimate (or choice), $\hat{\mathcal{H}}$, given some observation. For this derivation, assume that the observation is the (complex) spectrum $\mathbf{t}$.
The above leads to a simple decision theoretic formulation: Define a loss or cost matrix, $\mathbf{C}$, with elements $C_{\mathcal{H},\hat{\mathcal{H}}}$ that attaches a cost to each combination of $\mathcal{H}$ and $\hat{\mathcal{H}}$.

$$\mathbf{C} = \begin{pmatrix} C_{0,0} & C_{0,1} \\ C_{1,0} & C_{1,1} \end{pmatrix}$$

Typically, the cost should be low for a correct classification, and high for an incorrect one.

Expected cost

The expected costs of the two possible classifications are then:

$$\mathbb{E}\left(\mathrm{C}_{\mathcal{H},0} \mid \mathbf{t}\right) = \sum_{\mathcal{H}} \mathrm{C}_{\mathcal{H},0} \mathrm{P}\left(\mathcal{H} \mid \mathbf{t}\right)$$

$$\mathbb{E}\left(\mathrm{C}_{\mathcal{H},1} \mid \mathbf{t}\right) = \sum_{\mathcal{H}} \mathrm{C}_{\mathcal{H},1} \mathrm{P}\left(\mathcal{H} \mid \mathbf{t}\right)$$

Minimax

Now choose the classification, $\hat{\mathcal{H}}$, that has the smaller expected cost: Choose $\hat{\mathcal{H}} = 1$ if

$$\sum_{\mathcal{H}} C_{\mathcal{H},1} P(\mathcal{H} \mid \mathbf{t}) < \sum_{\mathcal{H}} C_{\mathcal{H},0} P(\mathcal{H} \mid \mathbf{t})$$

so,

$$\frac{P(\mathcal{H} = 1 \mid \mathbf{t})}{P(\mathcal{H} = 0 \mid \mathbf{t})} > \frac{C_{0,1} - C_{0,0}}{C_{1,0} - C_{1,1}}.$$

Bayesian test

Apply Bayes' theorm to the posterior terms. Choose $\hat{\mathcal{H}} = 1$ if

$$\underbrace{\frac{p\left(\mathbf{t} \mid \mathcal{H}=1\right)}{p\left(\mathbf{t} \mid \mathcal{H}=0\right)}}_{\text{Likelihood ratio, } L\left(\cdot\right)} \cdot \underbrace{\frac{P\left(\mathcal{H}=1\right)}{P\left(\mathcal{H}=0\right)}}_{\text{Prior ratio}} > \underbrace{\frac{C_{0,1}-C_{0,0}}{C_{1,0}-C_{1,1}}}_{\text{Cost ratio}}.$$

i.e., A likelihood ratio test.

Pragmatic test

Typically

- You don't know the prior ratio more accurately than "Ooh, about 1".
- You don't know the cost ratio.

But

- You can empirically guess a threshold (Ugh. . .)

So, choose $\hat{\mathcal{H}} = 1$ if

$$\frac{\mathrm{p}\left(\mathbf{t} \mid \mathcal{H} = 1\right)}{\mathrm{p}\left(\mathbf{t} \mid \mathcal{H} = 0\right)} > \lambda.$$

where λ is "just some number".

Two models

The Gaussian model can be used to do VAD.
We need two models:

- One for $\mathcal{H} = 0$, the noise model.
- One for $\mathcal{H} = 1$, the speech (plus noise) model.

One likelihood ratio

$$p\left(\mathbf{t} \mid \mathcal{H}=0\right)=\prod_{f=1}^{K} \frac{1}{\pi \nu_f} \exp\left(-\frac{|t_f|^2}{\nu_f}\right),$$

$$p\left(\mathbf{t} \mid \mathcal{H}=1\right)=\prod_{f=1}^{K} \frac{1}{\pi(\sigma_f+\nu_f)} \exp\left(-\frac{|t_f|^2}{\sigma_f+\nu_f}\right).$$

$$\frac{p\left(\mathbf{t} \mid \mathcal{H}=1\right)}{p\left(\mathbf{t} \mid \mathcal{H}=0\right)}=\prod_{f=1}^{K} \frac{\nu_f}{\sigma_f+\nu_f} \exp\left(\frac{\sigma_f}{\sigma_f+\nu_f} \cdot \frac{|t_f|^2}{\nu_f}\right).$$

Gaussian model problems

The Gaussian model

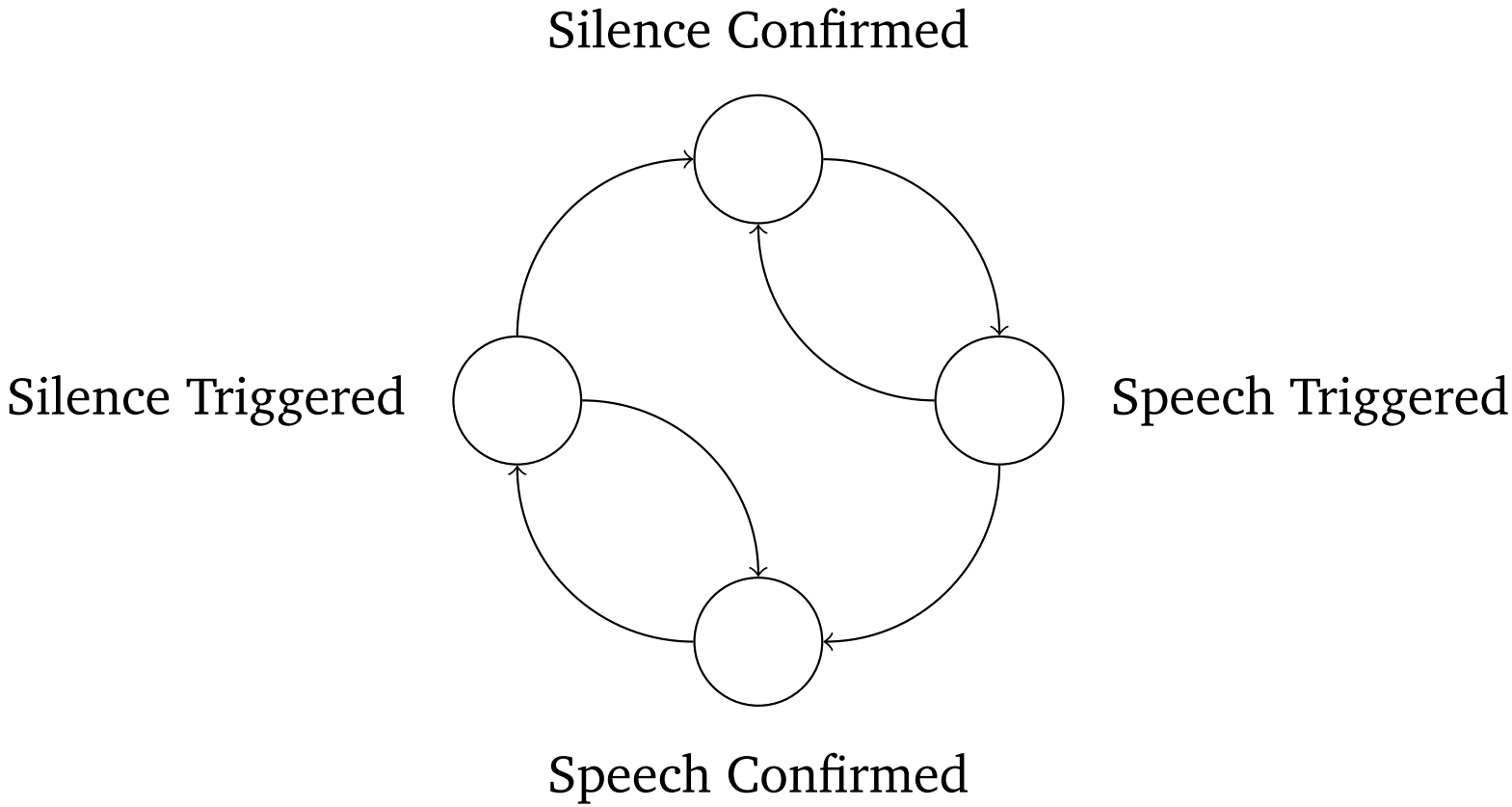
- Does not make any assumptions or predictions beyond one frame.
- Responds to anything that does not look like noise.
- It's sensitive to spikes.

It is insensitive to

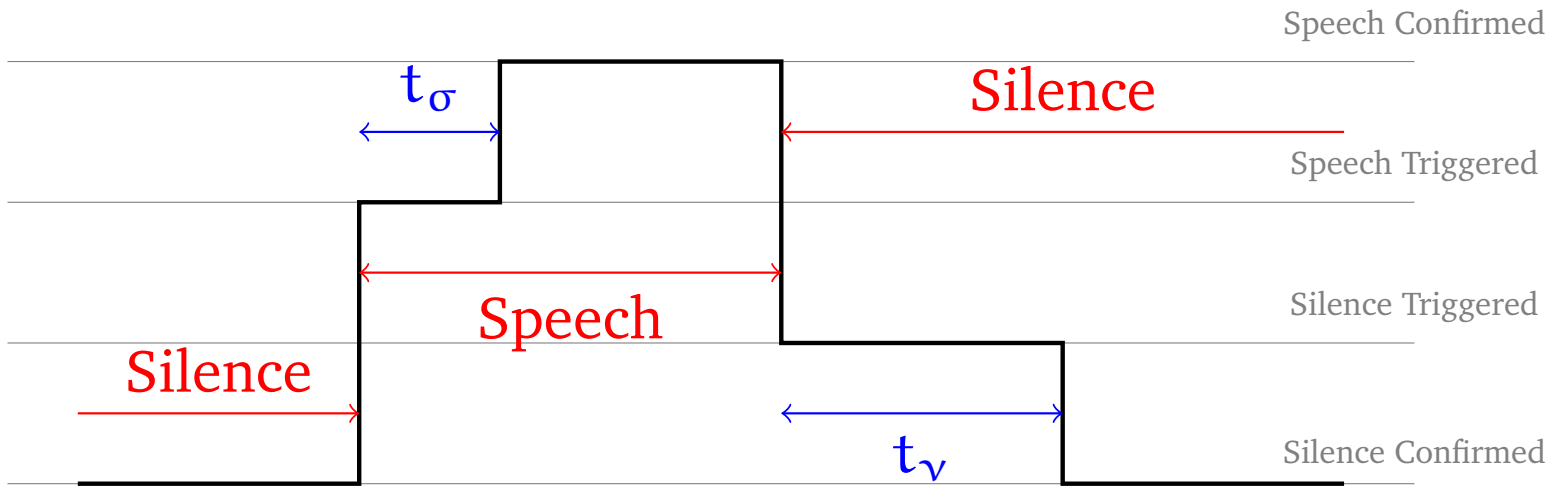
- Quiet utterances, especially utterances that become quiet.

In fact, these problems apply to any frame based or energy based model.

State machine



State machine levels



Minimum times:

- t_σ Minimum speech time.
- t_ν Minimum silence time.

Summary

VAD is one of the most subjective things in speech processing

- Everyone has their own algorithm.
- Many ad-hoc techniques exist.

You can do it in a principled way:

- Cost function based implementation.

...but the costs are somewhat subjective.