models

Definition

Speech is a continuous signal, which means that consecutive samples of the signal are correlated (see figure on the right). In particular, if we know a previous sample x_{n-1}, we can make a *prediction* of the current sample, $\hat{x}_n = x_{n-1}$, such that $\hat{x}_n \approx x_n$. By using more previous samples we have more information, which should help us make a better prediction. Specifically, we can define a predictor which uses M previous samples to predict the current sample x_n as

$$\hat{x}_n = -\sum_{k=1}^{M} a_k x_{n-k}.$$

This is a *linear predictor* because it takes a linearly weighted sum of past components to predict the current one.

The error of the prediction, also known as the *prediction residual* is

$$e_n = x_n - \hat{x}_n = x_n + \sum_{k=1}^{M} a_k x_{n-k} = \sum_{k=0}^{M} a_k x_{n-k},$$

where $a_0=1$. This explains why the definition $\hat{x}_n$ included a minus sign; when we calculate the residual, the double negative disappears and we can collate everything into one summation.

Vector notation

Using vector notation, we can make the expressions more compact

$$e = Xa$$

where

$$e = \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_{N-1} \end{bmatrix}, \quad X = \begin{bmatrix} x_0 & x_{-1} & \dots & x_M \\ x_1 & x_0 & \dots & x_{M-1} \\ \vdots & \vdots & & \vdots \\ x_{N-1} & x_{N-2} & \dots & x_{N-M} \end{bmatrix},$$

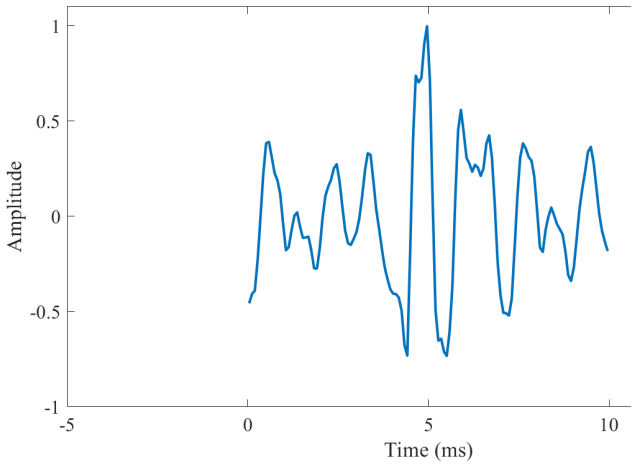
Here we calculated the residual for a length N frame of the signal.

Parameter estimation

Vector a holds the unknown coefficients of the predictor. To find the best possible predictor, we can minimize the minimum mean-square error (MMSE). The square error is the 2-norm of the residual, $\|e\|^2 = e^T e$. The mean of that error is defined as the expectation

$$E\left[\|e\|^2\right] = E\left[a^T X^T X a\right] = a^T E\left[X^T X\right] a = a^T R_x a,$$

A short segment of speech. Notice how consecutive samples are mostly near each other, which means that consecutive samples are correlated.



where $R_x = E\left[X^T X\right]$ and $E[\cdot]$ is the expectation operator. Note that, as shown in the autocorrelation

models

If we would directly minimize the mean-square error $E\left[\|e\|^2\right]$, then clearly we would obtain the trivial solution *a=0*, which is not particularly useful. However that solution contradicts with the requirement that the first coefficient is unity, a_0*=1*. In vector notation we can equivalently write

$$a_0 - 1 = u^T a - 1 = 0, \qquad \text{where}\, u = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The standard method for quadratic minimization with constraints is to use a Langrange multiplier, λ, such that the objective function is

$$\eta(a, \lambda) = a^T R_x a - 2\lambda\left(a^T u - 1\right).$$

This function can be heuristically interpreted such that λ is a free parameter. Since our objective is to minimize $a^T R_x a$ if $a^T u - 1$ is non-zero, then the objective function can become arbitrarily large. To allow any value for λ, the constraint must therefore be zero.

The objective function is then minimized by setting its derivative with respect to *a* to zero

$$0 = \frac{\partial}{\partial a}\eta(a, \lambda) = \frac{\partial}{\partial a}\left[a^T R_x a - 2\lambda\left(a^T u - 1\right)\right] = 2R_x$$

It follows that the optimal predictor coefficients are found by solving

$$R_x a = \lambda u.$$

Since R_x, is symmetric and Toeplitz, the above system of equations can be efficiently solved using the Levinson-Durbin algorithm with algorithmic complexity $O(M^2)$. However, note that with direct solution we obtain $a' := \frac{1}{\lambda}a = R_x^{-1}u$ that is, instead of *a* we get *a* scaled with λ. However, since we know that a_0*=1*, we can find *a* by $a = \lambda a' = \frac{a'}{a'_0}$.

Spectral properties

Linear prediction is usually used to predict the current sample of a time-domain signal x_n. The usefulness of linear prediction however becomes evident by studying its Fourier spectrum Specifically, since *e=Xa*, the corresponding Z-domain representation is

$$E(z) = X(z)A(z) \qquad \Rightarrow \qquad X(z) = \frac{E(z)}{A(z)},$$

where *E(z)*, *X(z)*, and *A(z)*, are the Z-transforms of e_n, x_n and a_n, respectively. The residual *E(z)* is white-noise, whereby the inverse $A(z)^{-1}$, must follow the shape of *X(z)*.

In other words, the linear predictor models the macro-shape or *envelope* of the spectrum.

models

order

Linear prediction has a surprising connection with physical modelling of speech production. Namely, a linear predictive model is equivalent with a *tube-model of the vocal tract* (see figure on the right). A useful consequence is that from the acoustic properties of such a tube-model, we can derive a relationship between the physical length of the vocal tract *L* and the number of parameters *M* of the corresponding linear predictor as

$$M = \frac{2 f_s L}{c},$$

where f_S is the sampling frequency and *c* is the speed of sound. With an air-temperature of 35 C, the speed of sound is *c*=350m/s. The mean length of vocal tracts for females and males are approximately 14.1 and 16.9 cm. We can then choose to overestimate *L*=0.17m. At a sampling frequency of 16kHz, this gives $M \approx 17$. The linear predictor will catch also features of the glottal oscillation and lip radiation, such that a useful approximation is $M \approx \mathrm{round}\left(1.25 \frac{f_s}{1000}\right)$. For different sampling rates we then get the number of parameters *M* as

f_S	M
8 kHz	10
12.8 kHz	16
16 kHz	20

Observe however that even if a tube-model is equivalent with a linear predictor, the relationship is non-linear and highly sensitive to small errors. Moreover, when estimating linear predictive models from speech, in addition to features of the vocal tract, we will also capture features of glottal oscillation and lip-radiation It is therefore very difficult to estimate meaningful tube-model parameters from speech. A related sub-field of speech analysis is glottal inverse filtering, which attempts to estimate the glottal source from the acoustic signal. A necessary step in such inverse filtering is to estimate the acoustic effect of the vocal tract, that is, it is necessary to estimate the tube model.

Uses in speech coding

Linear prediction has been highly influential especially in early speech coders. In fact, the dominant speech coding method is code-excited linear prediction (CELP), which is based on linear prediction.

Alternative representations (advanced topic)

Suppose scalars $a_{m,k}$, are the coefficients of an *M*th order linear predictor. Coefficients of consecutive orders *M* and *M*+1 are then related as

models

---

where the real valued scalar $\gamma_M \in (-1, +1)$ is the *M*th reflection coefficient. This formulation is the basis for the Levinson-Durbin algorithm which can be used to solve the linear predictive coefficients. In a physical sense, reflection coefficients describe the amount of the acoustic wave which is reflected back in each junction of the tube-model. In other words, there is a relationship between the *cross-sectional areas* S_k of each tube-segment and the reflection coefficients as

$$\gamma_k = \frac{S_k - S_{k+1}}{S_k + S_{k+1}}.$$

Furthermore, the logarithmic ratio of cross-sectional areas, also known as the *log-area ratios*, are defined as

$$A_k = \log \frac{S_k}{S_{k+1}} = \log \frac{1 - \gamma_k}{1 + \gamma_k}.$$

This form has been used in coding of linear predictive models, but is today mostly of historical interest.

无标签