Lecture 7: Aperiodicity Spectrum

Mark Hasegawa-Johnson

ECE 537: Speech Processing

Outline

Source-Filter Model of Speech

$$s[n] = h[n] * v[n]$$

- $v[n]$, the excitation function, is produced at the glottis for voiced speech, or at another constriction for unvoiced speech.
- Since the glottis is so much smaller than any other constriction, its behavior is not much affected by the shape of the rest of the vocal tract.
- The shape of the rest of the vocal tract determines $h[n]$, the vocal tract impulse response.

Models of Excitation

- One-mass model:

$$U_M(t) = h(t) * U_G(t)$$

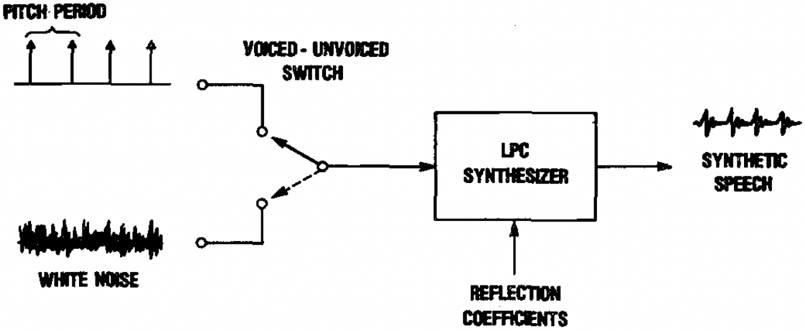
$$s(t) = |L_r| \frac{dU_M}{dt} \propto h(t) * v(t)$$

$$v(t) = \frac{dU_G}{dt}$$

- Vocoder model:

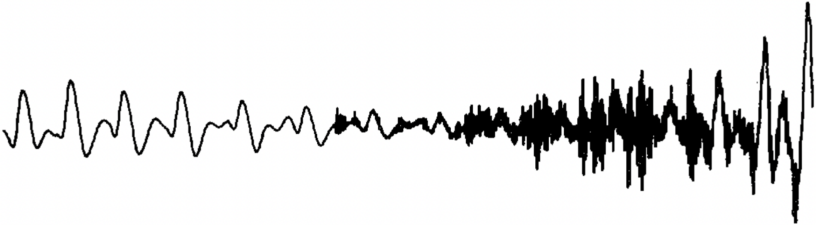
$$v[n] \begin{cases} = \sum_m \delta[n - mT_0] & s[n] \text{ voiced} \\ \sim \mathcal{N}(0,1) \text{ i.i.d.} & s[n] \text{ unvoiced} \end{cases}$$

The Vocoder Model



Atal and Remde, 1982, Fig. 1

The problem with the vocoder model: Real speech is mixed



Atal and Remde, 1982, Fig. 2

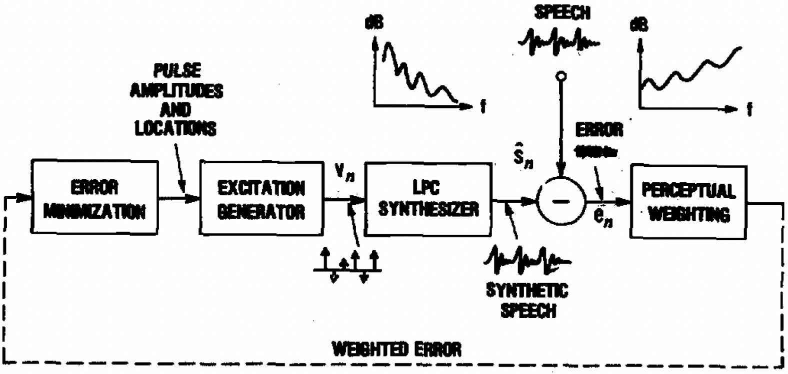
How do we fix the vocoder?

You can make synthetic speech sound better by adding white noise to your voiced excitation, with an amplitude about 10-20% the amplitude of the voiced excitation. But if you want a more controllable, automatic approach, you need one of two broad categories of approaches:

- **Add aperiodicity in the time domain:** Add extra impulses to your impulse train, timed and scaled to optimally match the target.
- **Add aperiodicity in the frequency domain:** Combine the DTFTs of a periodic source (impulse train) and an aperiodic source (white noise) with frequency-dependent interpolation parameters.

Outline

Analysis-by-Synthesis Multipulse Linear Predictive Coding



Analysis-by-Synthesis Method

$$v_K[n] = \sum_{k=1}^{K} a_k \delta[n - n_k]$$

$$\hat{s}_K[n] = \sum_{k=1}^{K} a_k h[n - n_k]$$

$$\mathcal{E}_K = \sum_{n} \left(s[n] - \hat{s}_K[n]\right)^2$$

The main idea: choose a_k and n_k to minimize $\mathcal{E}$. The article talks about perceptual weighting; we will ignore that until later in the course.

Analysis-by-Synthesis: Optimal amplitude

Suppose you know $n_1, \ldots, n_K$, and you know $a_1, \ldots, a_{K-1}$, but you don't yet know a_K. How can you find it?

$$\hat{s}_K[n] = \hat{s}_{K-1}[n] + a_K h[n - n_K]$$

$$\mathcal{E}_K = \sum_n \left(s[n] - \hat{s}_{K-1}[n] - a_K h[n - n_K]\right)^2$$

Differentiate the last line w.r.t. a_K and set it to zero to find that

$$a_{K,opt} = \frac{R_{EH}[n_K]}{R_{HH}[0]}$$

where

$$R_{EH}[n_k] = \sum_n \left(s[n] - \hat{s}_{k-1}[n]\right) h[n - n_k]$$

$$R_{HH}[0] = \sum_n h^2[n]$$

Analysis-by-Synthesis: Optimal position

Now we have the optimal value of a_K for each n_K, but we don't know n_K. How can we find it?

$$\begin{aligned}\mathcal{E}_{K,opt} &= \sum_n \left(s[n] - \hat{s}_{K-1}[n] - a_{K,opt}h[n-n_K]\right)^2 \\ &= \mathcal{E}_{K-1} - \frac{R_{EH}^2[n_K]}{R_{HH}[0]}\end{aligned}$$

So we only need to find the value of n_K that maximizes $R_{EH}[n_K]$.

Analysis-by-Synthesis: Optimal everything

Suppose you don't yet know anything. How can you proceed?

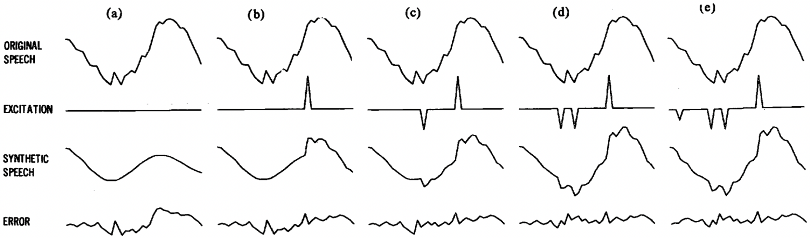
- **Initialize:** Set $\hat{s}_0[n]$ to be the tails of the impulse responses hanging over from the previous frame. Calculate the resulting $\mathcal{E}_0$.
- **Iterate:** For $k = 1, \ldots, K$: Choose

$$n_k = \text{argmax}\, R_{EH}[n_k]$$

$$a_k = \frac{R_{EH}[n_k]}{R_{HH}[0]}$$

- **Terminate:** End when the error is low enough, or when you can't afford to add any more impulses.

Analysis-by-Synthesis Visualized



Outline

Adding aperiodicity in the frequency domain

- Compute periodic and aperiodic excitations:

$$v_p[n] = \sum_m \delta[n - mT_0]$$
$$v_a[n] \sim \mathcal{N}(0,1) \text{ i.i.d}$$

- Take their DTFT:

$$V_p(\omega) = \mathrm{DTFT}\{v_p[n]\}$$
$$V_a(\omega) = \mathrm{DTFT}\{v_a[n]\}$$

- Add them with aperiodicity weights, where $\alpha_p^2(\omega) + \alpha_a^2(\omega) = 1$:

$$V(\omega) = \alpha_p(\omega) V_p(\omega) + \alpha_a(\omega) V_a(\omega)$$

The aperiodicity spectrum

Kawahara et al. (2008) suggest

$$\alpha_p^2(\omega) = \frac{\sigma_P^2(\omega)}{\sigma_P^2(\omega) + \sigma_N^2(\omega)}$$

- $\sigma_P^2(\omega)$ is the portion of $|V(\omega)|^2$ that is predictable from periodicity
- $\sigma_N^2(\omega)$ is the portion that is unpredictable

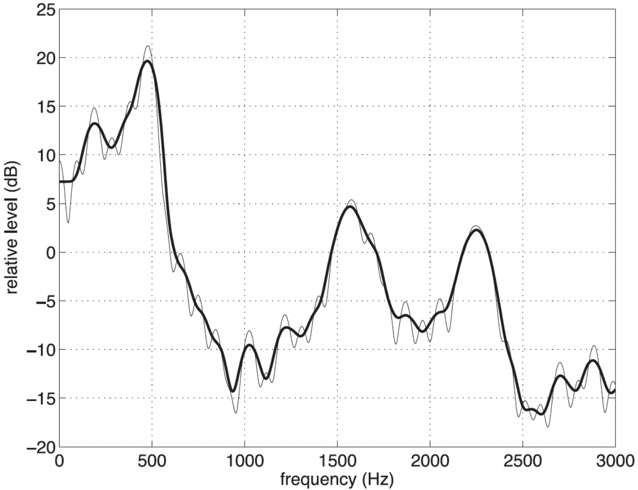
Aperiodicity spectrum: What is $|V(\omega)|^2$?

Remember $S(\omega) = H(\omega)V(\omega)$. Kawahara et al. use the following model:

- The speech power spectrum is the Tandem spectrum, $|S(\omega)|^2 \approx P_T(\omega)$.
- The vocal tract transfer function is the STRAIGHT spectrum, $|H(\omega)|^2 \approx P_{TST}(\omega)$.
- We can calculate the excitation power spectrum, and then subtract out its mean, resulting in the "fluctuation spectrum:"

$$P_C(\omega) \equiv \frac{P_T(\omega)}{P_{TST}(\omega)} - 1$$

Tandem and Tandem-STRAIGHT Spectra



Note: $P_{TST}(\omega)$ (dark) is a smoothed version of $P_T(\omega)$ (light), so $P_T(\omega)/P_{TST}(\omega) - 1$ contains only the fluctuations.

Aperiodicity spectrum: What portion of $|V(\omega)|^2$ is periodic?

The excitation spectrum is periodic to the extent that $P_C(\omega)$ varies sinusoidally with a period of ω_0. We can measure that by convolving, in frequency, with a matched filter that is exactly sinusoidal with a period of ω_0:

$$\sigma_P^2(\omega) \approx \left| h_N(\omega) * P_C(\omega) \right|^2$$
$$h_N(\omega) = w_N(\omega) \exp\left(2\pi j \omega / \omega_0\right)$$

$w_N(\omega)$ is a Hamming window with a length of $2N\omega_0$. $N \approx 8$ is chosen experimentally to trade off frequency selectivity versus noise.

What portion of $|V(\omega)|^2$ is aperiodic?

How do we find the aperiodic portion, $\sigma_N^2(\omega)$? Kawahara et al. choose, instead, to set

$$\sigma_P^2 + \sigma_N^2 = \sigma_{MAX}^2$$

- Recall that $P_C(\omega) = \frac{P_T(\omega)}{P_{TST}(\omega)} - 1$ is normalized, so it doesn't depend on the signal energy.
- Instead, $P_C(\omega)$ only depends on the periodicity of the signal.
- Kawahara et al. create an ideal impulse train, then pass it through all of the above processing to find

$$\sigma_P^2(\omega) = \sigma_{MAX}^2 \quad \text{if} \quad s[n] = \sum_m \delta[n - mT_0]$$

  They note that this value of σ_{MAX}^2 only depends on the window used for the original STFT; for a normalized Hamming window, it is $\sigma_{MAX}^2 = 1/16$.

Aperiodicity spectrum: Putting it all together

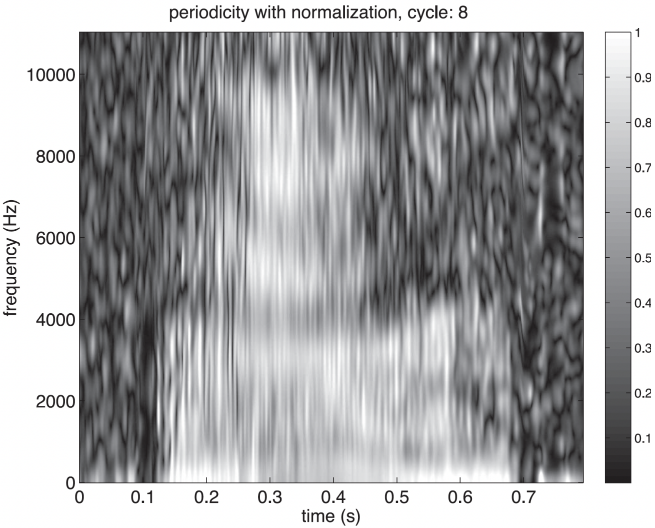
$$P_C(\omega) = \frac{P_T(\omega)}{P_{TST}(\omega)} - 1$$

$$\sigma_P^2(\omega) = |h_N(\omega) * P_C(\omega)|^2$$

$$\alpha_p^2(\omega) = \frac{\sigma_P^2(\omega)}{\sigma_{MAX}^2}$$

$$\alpha_a^2(\omega) = 1 - \alpha_p^2(\omega)$$

Fraction of excitation that is periodic: example



Outline

Summary

- Time-domain modeling of aperiodicity: MPLPC

$$v_K[n] = \sum_{k=1}^{K} a_k \delta[n - n_k]$$

$$\hat{s}_K[n] = \sum_{k=1}^{K} a_k h[n - n_k]$$

- Frequency-domain modeling of aperiodicity:

$$V(\omega) = \alpha_p(\omega) V_p(\omega) + \alpha_a(\omega) V_a(\omega)$$

$$\alpha_p(\omega) = \sqrt{\frac{\sigma_P^2(\omega)}{\sigma_{MAX}^2}}$$