Lecture 5: STRAIGHT — Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum

Mark Hasegawa-Johnson

ECE 537: Speech Processing

Outline

Review: STFT

$$
\begin{aligned}
X_n(\hat{\omega}) &\equiv \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\hat{\omega}m} \\
&= e^{-j\hat{\omega}(n-N+1)} \sum_{m'=0}^{N-1} w[(N-1)-m']x[n+m']e{-j\hat{\omega}m'} \\
&= w[n] * \left(x[n]e^{-j\hat{\omega}n}\right) \\
&= e^{-j\hat{\omega}n}\left(h_{\hat{\omega}}[n] * x[n]\right)
\end{aligned}
$$

$$
x[m] = \frac{w[n-m]x[m]}{w[n-m]} = \frac{\sum_{n=-\infty}^{\infty} w[n-m]x[m]}{\sum_{n=-\infty}^{\infty} w[n-m]}
$$

Review: DTFT of Windowed Periodic Signals

- The Fourier transform of an impulse train is an impulse train:

$$\sum_{m=-\infty}^{\infty} \delta[n-mT_0] \leftrightarrow \frac{2\pi}{T_0}\sum_{k=0}^{T_0-1} \delta\left(\omega-\frac{2\pi k}{T_0}\right)$$

- Voiced speech is created by sending a periodic voice source through a resonant pipe:

$$x[n]=w[n]\left(h[n]*\sum_{m=-\infty}^{\infty}\delta[n-mT_0]\right)$$

$$X(\omega)=\frac{1}{T_0}\sum_{k=0}^{T_0-1} H\left(\frac{2\pi k}{T_0}\right) W\left(\omega-\frac{2\pi k}{T_0}\right)$$

Putting them together: STFT of a Periodic Signal

$$x[m]w[n-m] = \left(h[m] * \sum_{\ell=-\infty}^{\infty} \delta[m-\ell T_0]\right) w[n-m]$$
$$= \left(\sum_{\ell=-\infty}^{\infty} h[m-\ell T_0]\right) w[n-m]$$

$$X_n(\omega) = \frac{1}{2\pi} X(\omega) * e^{-j\omega n} W(-\omega)$$
$$= \frac{1}{T_0} \sum_{k=0}^{T_0-1} H\left(\frac{2\pi k}{T_0}\right) e^{-j\left(\omega - \frac{2\pi k}{T_0}\right)n} W\left(\frac{2\pi k}{T_0} - \omega\right)$$

More Review: Symmetric Windows

In their 2008 article, Kawahara et al. assume that $W(\omega)$ is real-valued, which requires that $w[n]$ is a symmetric function of time, nonzero between $-\frac{N-1}{2} \leq n \leq \frac{N-1}{2}$. This assumption defines the timing of the windows: Each window is timed at its center, not at its start or end. Remember that

$$\begin{aligned}
w[n] \text{ real} &\leftrightarrow W(\omega) = W^*(-\omega) \\
w[n] \text{ real \& symmetric} &\leftrightarrow W(\omega) \text{ real \& symmetric} \\
w[n] \text{ real \& antisymmetric} &\leftrightarrow W(\omega) \text{ imaginary \& antisymmetric} \\
w[n] \text{ imaginary \& symmetric} &\leftrightarrow W(\omega) \text{ imaginary \& symmetric} \\
w[n] \text{ imaginary \& antisymmetric} &\leftrightarrow W(\omega) \text{ real \& antisymmetric}
\end{aligned}$$

In practice, we usually only use the first two properties.

Outline

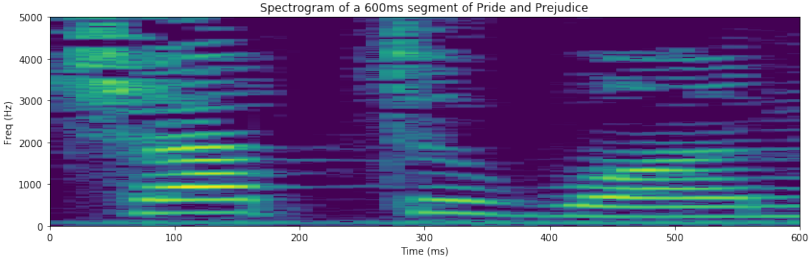
Definition of Spectrogram

$$S(n,\omega) = 20\log_{10}|X_n(\omega)| = 10\log_{10}|X_n(\omega)|^2$$

- Spectrogram = log periodogram. Periodogram is a (noisy) estimate of the power spectrum, so spectrogram is useful for both random and deterministic signals.
- One "Bell" ($\log_{10}|X_n(\omega)|^2$) is the amount of power lost by a telephone signal in one mile of copper cable.
- Most humans can hear a **percentage difference** in loudness corresponding to about one decibel ($10\log_{10}|X_n(\omega)|^2$).

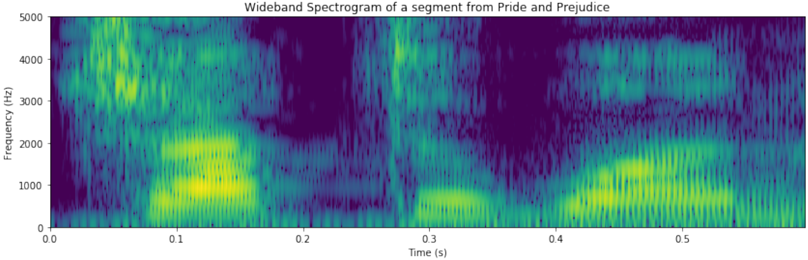
Narrowband Spectrogram

If the window length is greater than $2T_0$, the spectrogram is a quasiperiodic function of frequency, with period $\omega_0 = \frac{2\pi}{T_0}$.



Wideband Spectrogram

If the window length is less than T_0, the spectrogram is a quasiperiodic function of time, with period T_0.



How can we estimate $H(\omega)$?

$$X_n(\omega) = \frac{1}{T_0}\sum_{k=0}^{T_0-1} H\left(\frac{2\pi k}{T_0}\right) e^{-j\left(\omega-\frac{2\pi k}{T_0}\right)n} W\left(\frac{2\pi k}{T_0}-\omega\right)$$

- At the k^{th} harmonic, $|X_n(k\omega_0)| = \frac{W(0)}{T_0}|H(k\omega_0)|$.
- What is the most reasonable estimate of $H(\omega)$ between harmonics, and as a function of time between pitch pulses?

STRAIGHT and Tandem-STRAIGHT: Overview

What is the most reasonable estimate of $H(\omega)$ between harmonics, and as a function of time between pitch pulses?

- Classic-STRAIGHT (Kawahara, 1997): Linearly interpolate the spectrogram in time and frequency.
- Tandem-STRAIGHT (Kawahara et al., 2008): Average two samples of $|X_n(\omega)|^2$ in the time domain, then smooth its logarithm in the frequency domain using a piece-wise constant averager.

Outline

Speech Representation and Transformation using Adaptive Interpolation of weiGHTed spectrum

The original STRAIGHT algorithm smoothed $X_n(\omega)$ as:

$$S(\omega, n) = \sqrt{g^{-1}\left(h(\omega, n) ** g\left(|X_n(\omega)|^2\right)\right)}$$

- $**$ means 2D convolution (in **both** time and frequency)
- $g(\cdot)$ is a nonlinearity.
  - If you want to smooth the spectrogram directly, you use $g(\cdot) = \log(\cdot)$.
  - Kawahara suggests using $g(\cdot) = (\cdot)^{1/3}$.
- $h(\omega, n)$ is a bilinear interpolator—see next slide.

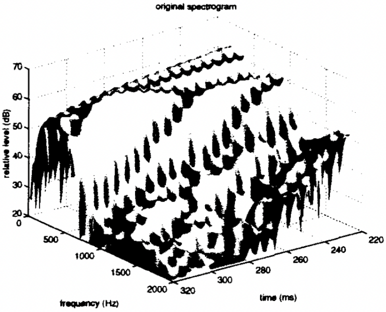
Bilinear Interpolation

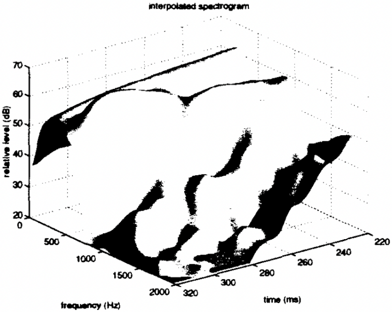
Bilinear interpolation is an idea from image processing. The bilinear interpolation kernel is a pyramid in the time-frequency plane:

$$h(\omega, n) = \frac{1}{4}\left(1 - \left|\frac{\omega}{\omega_0}\right|\right)\left(1 - \left|\frac{n}{T_0}\right|\right)$$

- If the input spectrogram consists of impulses at integer multiples of (ω_0, T_0), then convolution with $h(\omega, n)$ linearly interpolates between the impulses. (IMO this is the original motivation for the acronym "STRAIGHT" – bilinear interpolation creates a spectral envelope that almost a piece-wise linear function of time and frequency).
- A real-world spectrogram is smoothed in a similar way.

STRAIGHT Examples from Kawahara, 1997





What have I not told you?

- Good results with STRAIGHT depend on having excellent estimates of T_0 as a function of time.
- The original paper includes a detailed algorithm for pitch estimation, which we won't talk about today.

Outline

Tandem-STRAIGHT: Separate the time and frequency analyses

In their 2008 paper, Kawahara et al. reduced the computational complexity of STRAIGHT by quite a bit using a two-step method:

- To smooth in time, an assumption about the window spectrum permits the averaging of just two time points.
- To smooth in frequency, use a rectangular averager, then compensate for its overblurring of the spectrogram by subtracting sidelobes.

The Tandem spectogram: Two-point averaging in time

Recall that $X_n(\hat{\omega}) = w[n] * \left(x[n]e^{-j\hat{\omega}n}\right)$ is a lowpass function of time, with bandlimits given by the bandlimits of $W(\omega)$. That being the case, Kawahara et al. suggest that each band contains at most two harmonics of the fundamental. Suppose that one of these downmixes to $\omega = k\omega_0 - \hat{\omega}$, the other to $\omega + \omega_0$. Up to scalar multipliers, $X_n(\hat{\omega})$ and its squared magnitude are:

$$
\begin{aligned}
X_n(\hat{\omega}) &= A_k e^{j\theta_k} W(\omega) e^{j\omega n} + A_{k+1} e^{j\theta_{k+1}} W(\omega+\omega_0) e^{j(\omega+\omega_0)n} \\
|X_n(\hat{\omega})|^2 &= A_k^2 W^2(\omega) + A_{k+1}^2 W^2(\omega+\omega_0) \\
&\quad + 2A_k A_{k+1} W(\omega) W(\omega+\omega_0) \cos\left(\omega_0 n + \theta_{k+1} - \theta_k\right)
\end{aligned}
$$

Therefore $P_T(\hat{\omega}) = \frac{1}{2}\left(|X_n(\hat{\omega})|^2 + |X_{n+\frac{T_0}{2}}(\hat{\omega})|^2\right)$ has no temporal modulation, instead it only keeps the spectral envelope:

$$
P_T(\hat{\omega}) = A_k^2 W^2(\omega) + A_{k+1}^2 W^2(\omega+\omega_0)
$$

The oversmoothed spectrogram: Smooth in frequency with a rectangle

To smooth between harmonics, we can convolve in frequency by a rectangle with a width of ω_0. A computationally efficient way to do so is to calculate the cumulative spectrum $C(\omega)$, then difference it:

$$C(\omega) = \int_0^{\omega} P_T(\lambda) d\lambda$$

$$P_S(\omega) = C\left(\omega + \frac{\omega_0}{2}\right) - C\left(\omega - \frac{\omega_0}{2}\right)$$

The Tandem-STRAIGHT spectrogram: Compensate for oversmoothing

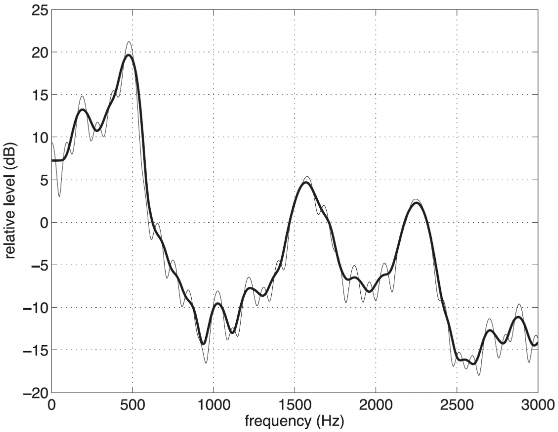
Here, OSS stands for "oversmoothed spectrogram." Smoothing by a rectangle is actually too much smoothing. Kawahara et al. compensate using a theorem from sampling theory to show that the spectral envelope can be reconstructed from its samples (one sample per ω_0) by convolving it with a discrete filter $q(\omega)$. Empirically, they find that the best results are obtained by applying the filter in log-spectral domain, thus

$$P_{TST}(\omega) = \exp\left(q(\omega) * \ln P_S\left(\frac{2\pi k}{T_0}\right)\right)$$

Where the compensating filter, $q(\omega)$, subtracts small portions of the neighboring harmonics:

$$q(\omega) = -0.047\delta(\omega + \omega_0) + \delta(\omega) - 0.047\delta(\omega - \omega_0)$$

Tandem and Tandem-STRAIGHT spectra (Kawahara et al., 2008)



The purpose for which STRAIGHT excels: Speech modification

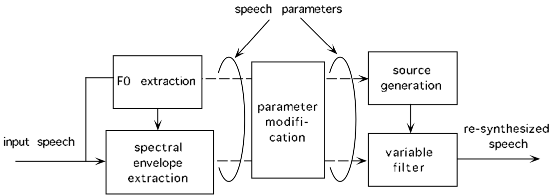


Figure 16 from Kawahara et al. (1998), "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds"

Outline

Summary

$$P_{ST}(n,\omega) = f^{-1}\left(h_2(\omega) * g^{-1}\left(h_1[n] * g\left(f\left(|X_n(\omega)|^2\right)\right)\right)\right)$$

where h_1 and h_2 are smoothing kernels in time and frequency, respectively, and where g and f are scalar nonlinearities.

- Classic-STRAIGHT: $f(\cdot) = (\cdot)^{1/3}, g(\cdot) = \cdot$, h_1 and h_2 are triangles with widths of $2T_0$ and $2\omega_0$.
- Tandem-STRAIGHT: $f(\cdot) = \ln(\cdot)$, $g(\cdot) = \exp(\cdot)$, h_1 is a pair of impulses separated by $\frac{T_0}{2}$, and h_2 is piece-wise-rectangular.