Lip Volume Velocity of Synthetic Speech

# Linear Model of Speech Production

The linear model for speech production describes a speech signal as the output of a linear system which, for voiced speech (vowels), is excited by an impulse train input. The impulse train signal is amplified by a scale factor , which depends on the intensity, or loudness, of the speech. The signal then passes through the glottal filter to produce the glottal volume velocity . The glottal filter is a critically damped second order low pass filter with one repeated real pole,

|  |  |
| --- | --- |
|  | (1) |

where is the cutoff frequency in rad/s and is the sampling period.

The glottal volume velocity then passes through the vocal tract filter to produce the lip volume velocity . The vocal tract can be modelled as a cascade of overdamped second order low pass filters, where each filter resonates at a particular formant frequency,

|  |  |
| --- | --- |
|  | (2) |

where is the maximum number of formants that can be captured by the model, is the -th formant frequency and is the 3dB bandwidth of the -th formant.

The lip volume velocity then passes through the lip radiation filter to produce the speech signal . The effect of lip radiation is a 6dB/octave lift in the spectrum of the signal, which can be modelled as a first order differentiator,

|  |  |
| --- | --- |
|  | (3) |

The spectrum of the resulting speech signal is given by

|  |  |
| --- | --- |
|  | (4) |

where is the transfer function of the entire speech production model, capturing the action of the glottis, vocal tract and lip radiation,

|  |  |
| --- | --- |
|  | (5) |

# Lip Volume Velocities

The lip volume velocity can be obtained by taking the running total (integral) of the speech signal,

|  |  |
| --- | --- |
|  | (6) |

In the z-domain, integration can be described as a linear system with a pole at ,

|  |  |
| --- | --- |
|  | (7) |

The transfer function relating the lip volume velocity to the impulse train input is

|  |  |
| --- | --- |
|  | (8) |

which is stable, since all poles lie within the unit circle. Note that the poles of the system are of the form

|  |  |
| --- | --- |
|  | (9) |

This means that a bounded produces a bounded , which is expected.

Synthetic speech tends to produce unbounded lip volume velocities, which indicates that it does not originate from the source-filter model. It is well known that integration is an unstable process, since the pole at lies on the unit circle in the z-plane. This means that a bounded input speech signal can potentially produce an unbounded output lip volume velocity when the integral is taken.

# Impact of the Microphone

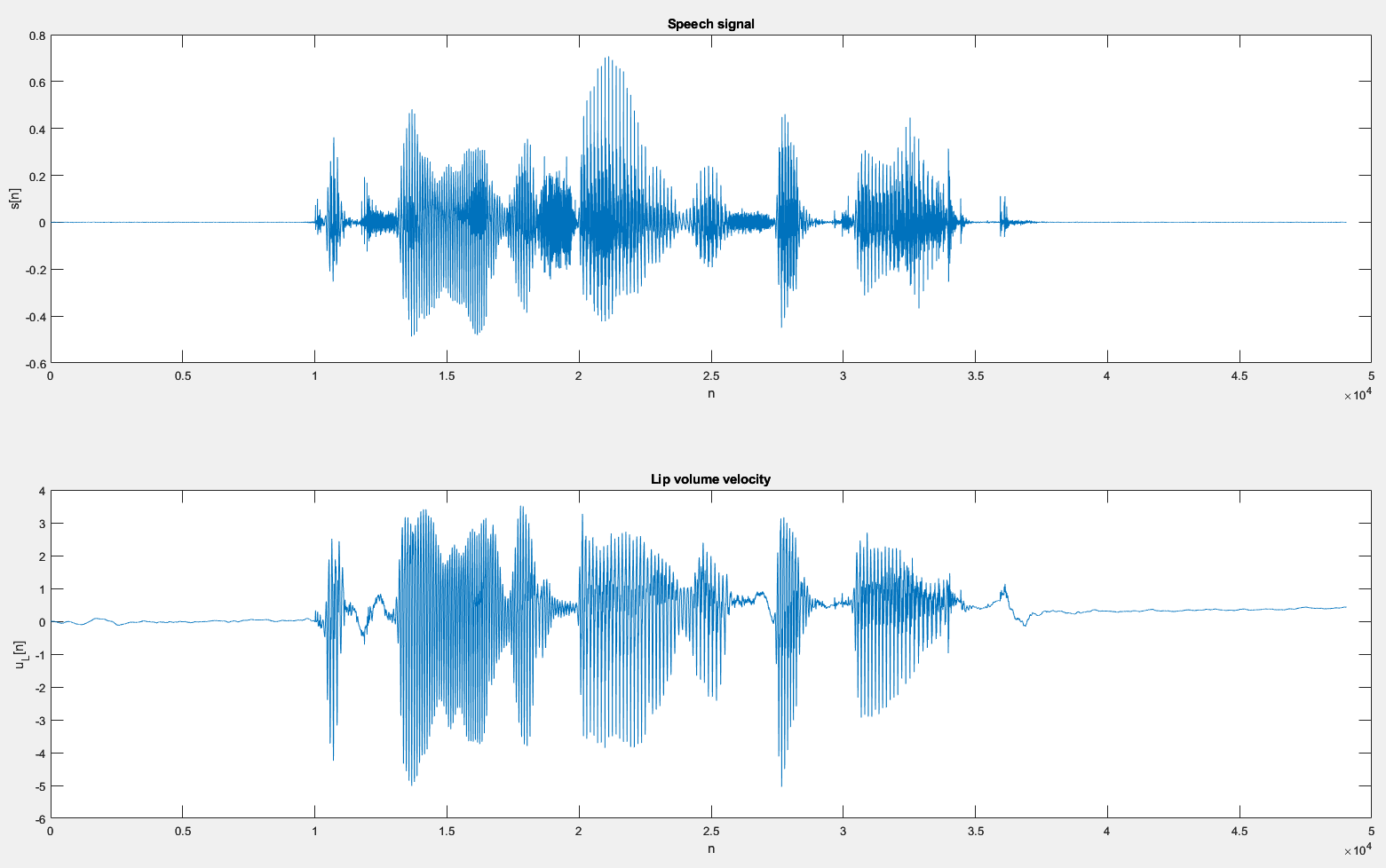
When speech is recorded by the microphone, frequencies within 0 – 20Hz are not captured, thus the recorded speech has zero mean, resulting in a lip volume velocity which is also zero mean. The synthetic speech models appear to contain frequencies within the 0 – 20Hz range, causing a variable DC offset to the lip volume velocity. Note that, in the z-domain, the pole at causes a spike in the frequency response at DC,

|  |  |
| --- | --- |
|  | (10) |

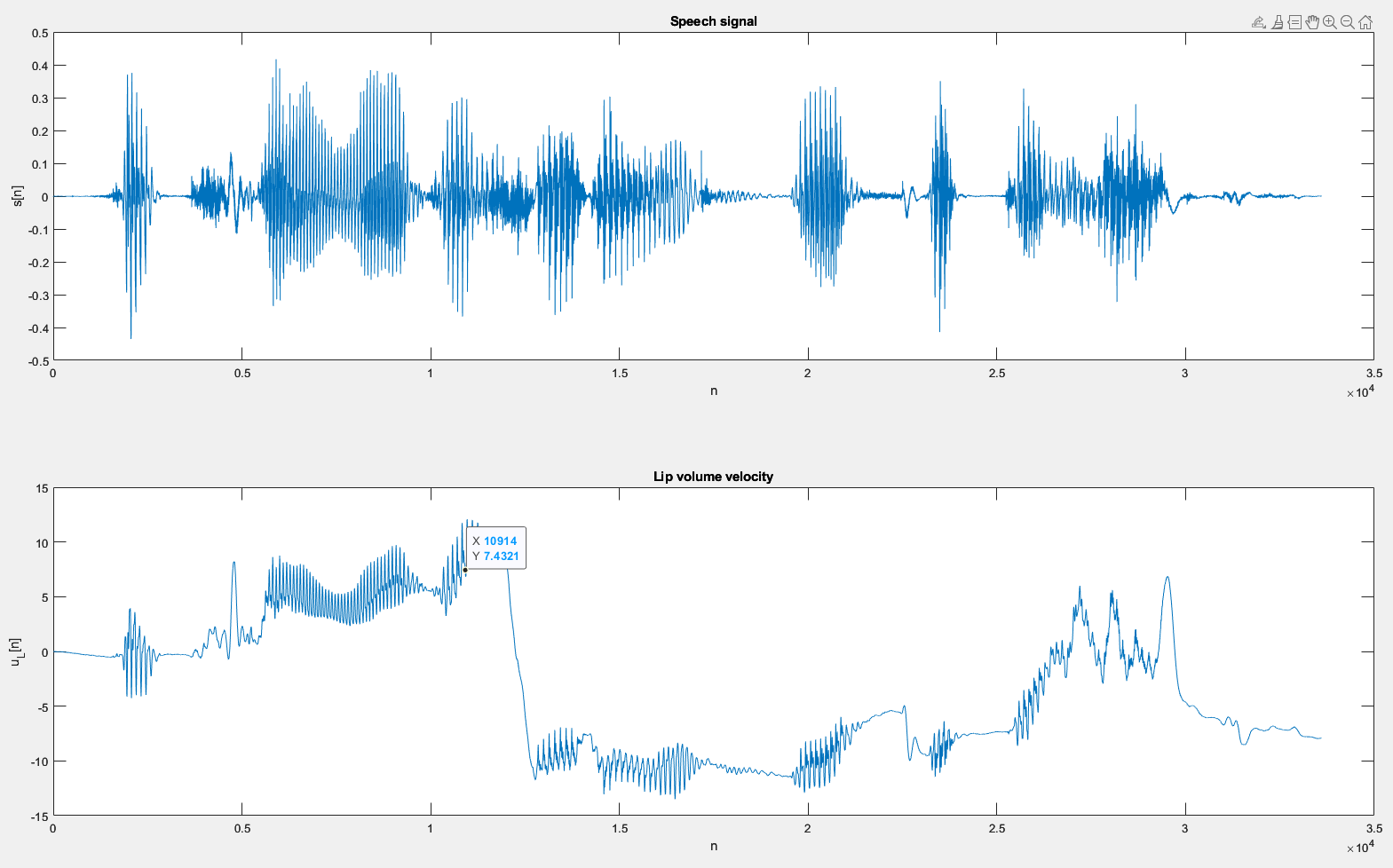
which causes low frequencies to be amplified and high frequencies to be attenuated.

“The key you designed will fit the lock”

**Natural Speech - Male**

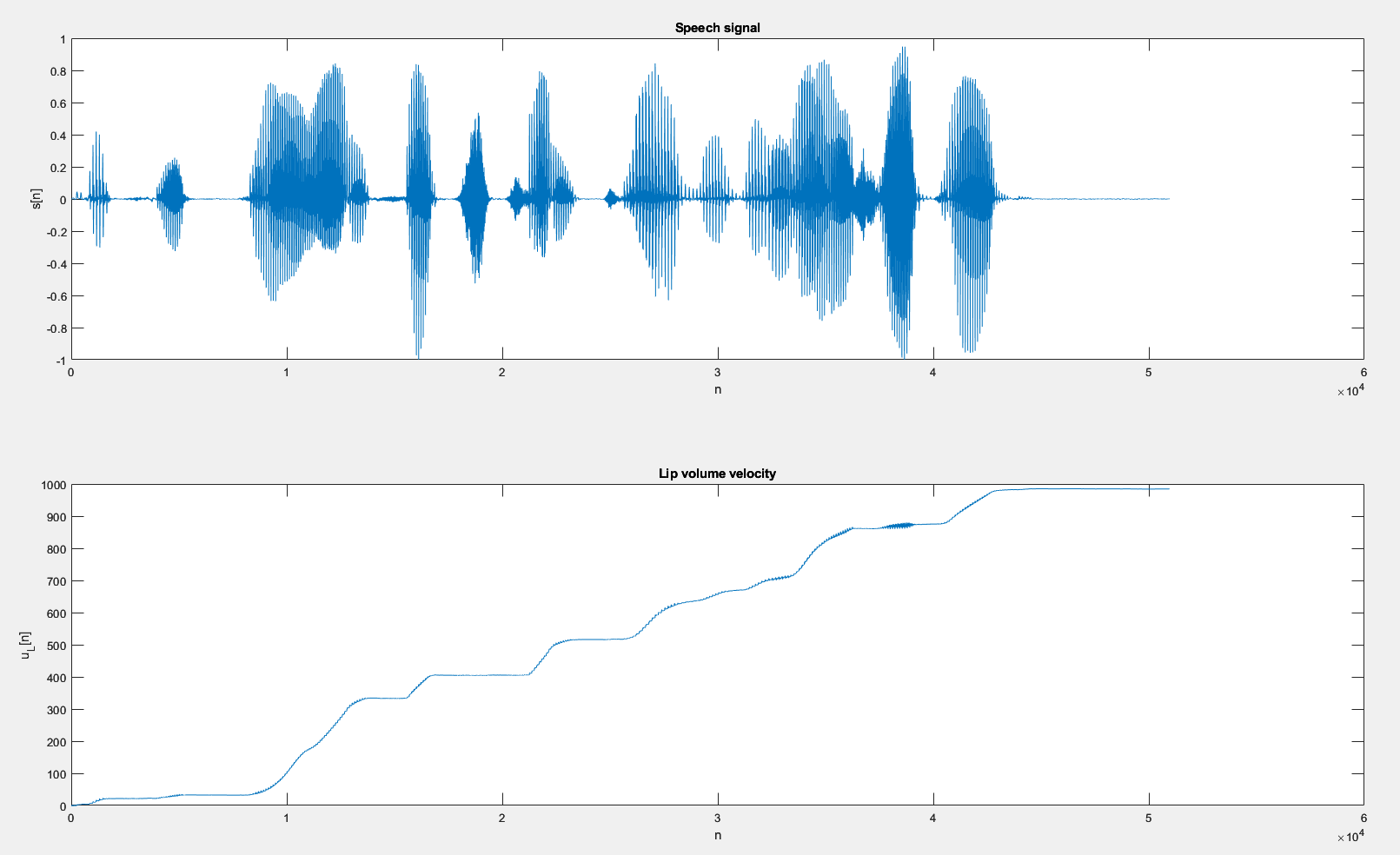
****

**Tacotron 2 – Male**

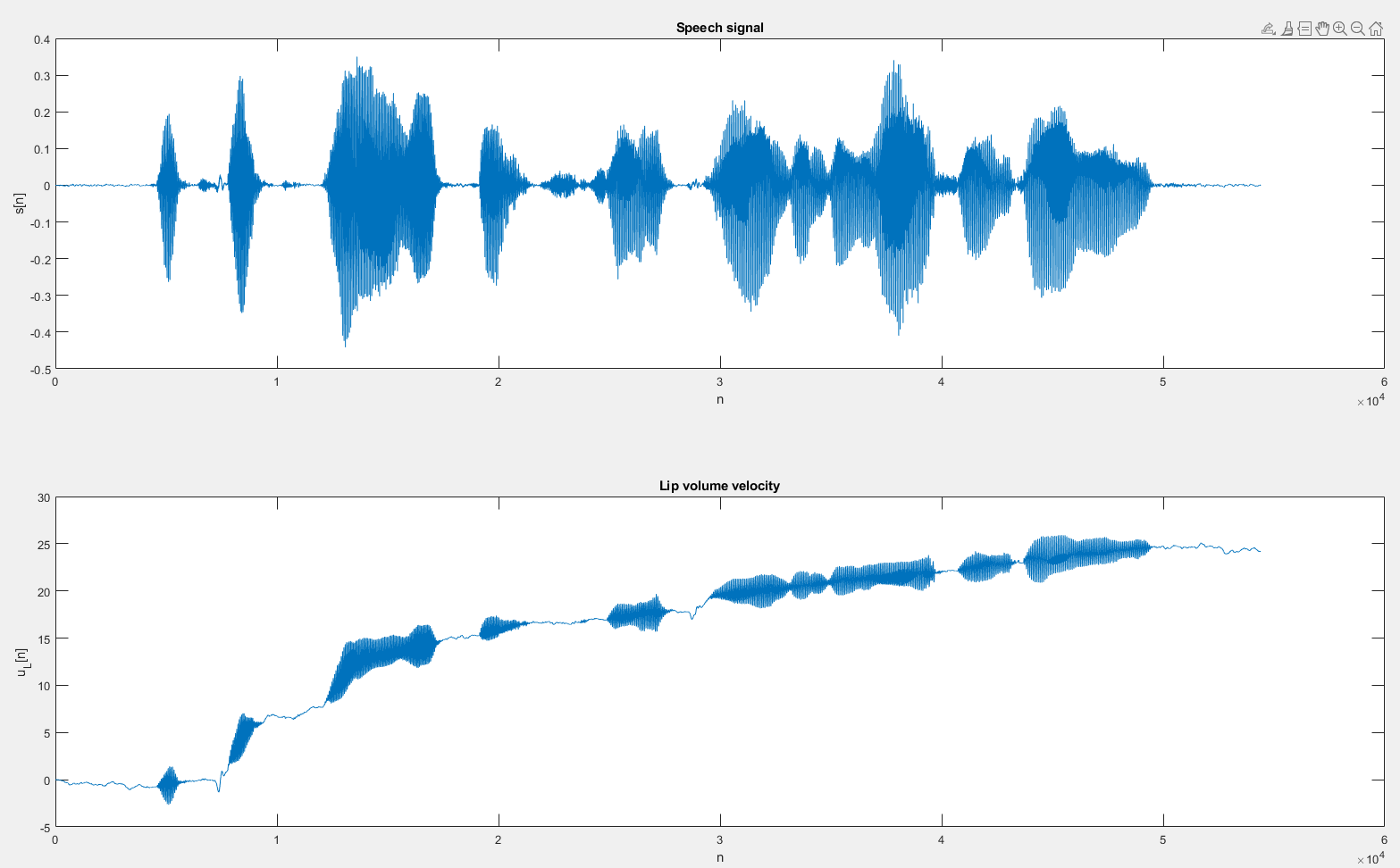
****

“The quick brown fox jumped over the lazy dog”

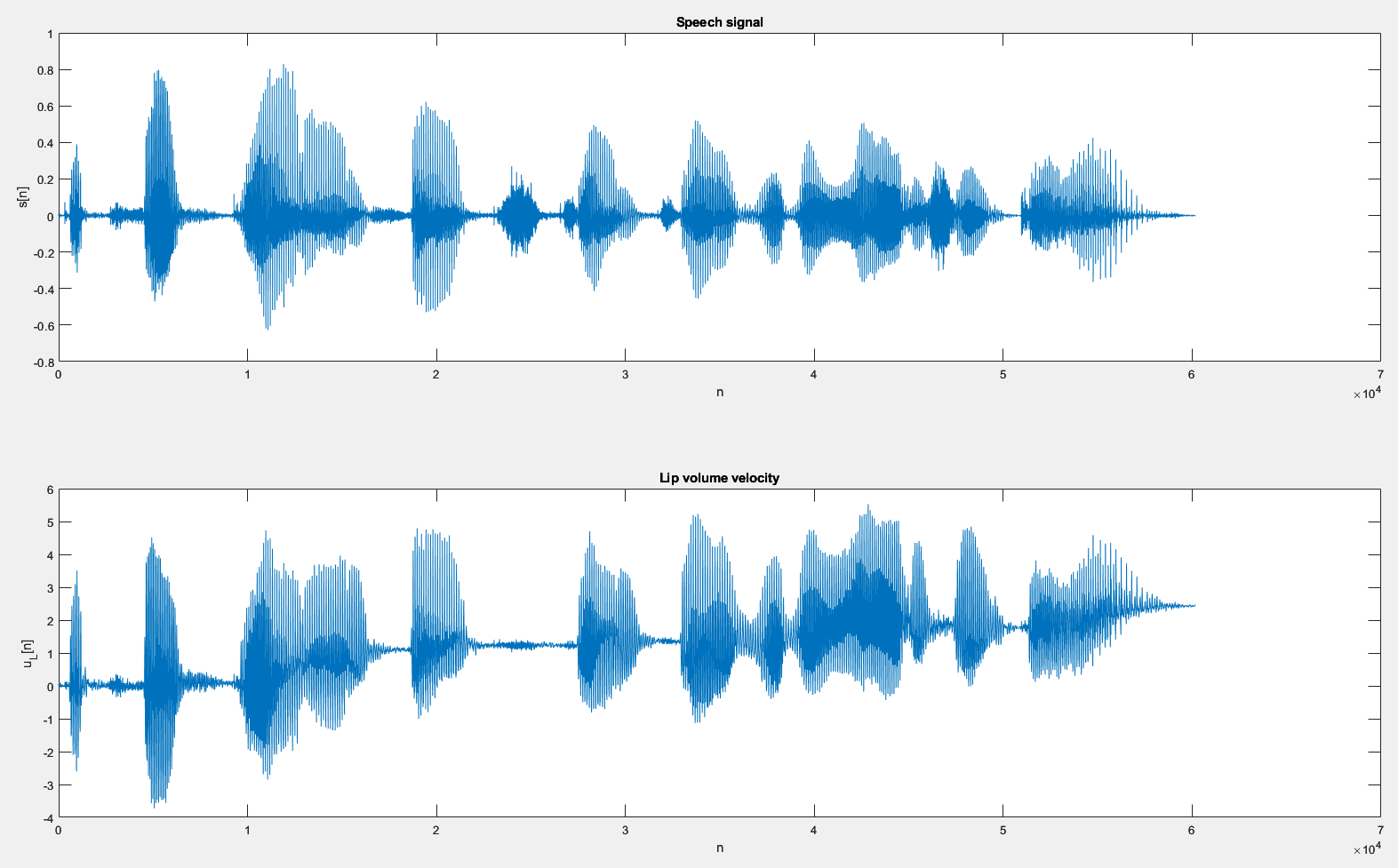
**MaryTTS - Male**

****

**Festival - Female**

****

**FastSpeech2 - Female (American)**

****

* Natural speech has a lip volume velocity which is approximately zero mean
* Tacotron 2 speech lip volume velocity mean value is seemingly random
* MaryTTS, Festival and Fastspeech2 have lip volume velocities which have increasing mean values. MaryTTS has the biggest increase, followed by Festival and then Fastspeech2.