What Are Acoustic Landmarks, and What Do They Describe?

Aug. 15, 2016
Dr. Joel MacAuslan

In speech acoustics, *landmarks* are patterns that mark certain speech-production events. Speech-acoustic landmarks come in two classes: *peak* and *abrupt*.

Peak: At present, the peak landmarks detected in SpeechMark® are *vowel landmarks* (VLMs) and *frication landmarks*. These are identified as instants in an utterance at which a maximum (or peak) of harmonic power or of fractal dimension occurs, respectively, and may be considered the centers of the vowels or fricated intervals (resp.). When plotted with SpeechMark functions, they are drawn below the waveform, labeled by uppercase letters: V or F. Frication landmarks are more fully described elsewhere (e.g., "Frication Peak Landmarks" on the SpeechMark website) and will be ignored here.[1]

Abrupt: *Abrupt* or *abrupt-consonantal landmarks* (AC LMs, or simply LMs) have a more complex specification.

It is helpful first to distinguish laryngeal-source from vocal-tract events. We denote the former by "+g" (overall onset) or "-g" (overall offset), by "+p" (onset of periodicity) or "-p" (offset, likewise), or by "+j" (upward jump of fundamental frequency, F_0) or "-j" (downward, likewise). The detailed rule for the critical $\pm$g is particularly complex. However, the central observation is easily stated:
Vocal-tract excitation by the laryngeal source is characterized by well-developed voicing.

> Voicing is considered *well developed* when there is evidence of sustained periodic excitation of at least minimal amplitude, as measured over intervals of several milliseconds.[2] In spectrogram terms: A narrow-band spectrogram shows clearly defined, smooth, approximately horizontal stripes, reflecting the harmonics of the excitation signal. The spacing between stripes defines the fundamental frequency. Apart from occasional jumps ($\pm$j), this frequency must lie within a range specified by the user, or by the client software, or by default. (The current defaults for human speech are: maximum F_0 = 350 Hz, minimum F_0 = 1/5 of maximum; these are typical of adults, especially females.) The limits of such an interval are denoted by +p and –p events.
>
> Additionally, voicing is considered to be present in a segment of the signal if it occurs shortly before a segment with well developed voicing with (a) similar power, and (b) similar spectral slope to the well-voiced segment. Currently, "shortly before" is up to 50 ms. Such a segment reflects glottalization or other irregular laryngeal motion.
>
> Both "g" and "p" LMs occur only in pairs. (Jumps do not.) So we may speak of voicing or of periodic voicing as an attribute of an entire *segment* of a signal, i.e., of the interval between +g and –g, or between +p and –p, similarly.

Thus, a +g/-g interval must include at least one +p/-p subinterval. However, it may contain more than one, and it may contain both F_0 jumps and intervals of irregular motion between +p/-p subintervals. Sometimes it may contain +p/-p intervals separated only by jumps, either upward or

---

[1] Peak-type F landmarks occur at the centers of fricated *intervals*. They should not be confused with the "f" and "v" landmarks (described below), which identify abrupt *events*, often at the onset and offset of frication.

[2] Specifically, the evaluation currently occurs every 8 ms, using data over a small multiple of this length. By default, the multiple is 4, so identified voicing typically lasts for at least 32 ms.

downward. Many voiced intervals start with periodicity, so for these intervals, +g and +p are coincident; and similarly for coincident –g and –p LMs.

Informally, but very usefully, the remaining LMs are identified as instants at which the signal shows evidence of *rapid change across multiple frequency ranges, on multiple time scales*.

In each case, AC LMs are classified as *onset* (+) or *offset* (-) type. They are also classified as *voiced* or *unvoiced*, according to their location in a voiced segment (between +g and –g) or an unvoiced one.

Processing begins by computing the power in each of several frequency bands. At present, the SpeechMark system normally uses five bands, from 800 to 8000 Hz for adults, or 1200 to 8000 Hz for infants. The instantaneous power is smoothed over two time scales, approximately 25 ms ("fine") and 50 ms ("coarse"): Coarse smoothing suppresses too-brief events, fine smoothing allows higher-precision placement.

> ***A landmark is detected if power rises or falls by 6 dB simultaneously at <u>both</u> fine and coarse time scales, and in at least 3 of the 5 bands.***

In practice, simultaneity is measured to a precision of 20 ms. That is, three bands must show 6-dB increases or decreases within 20 ms of each other in the coarsely smoothed power contours, *and* three must show the same in the finely smoothed contours, *and* the coarse and fine increases or decreases must lie within 20 ms of each other.

In the simplest case, power rises in all the bands, on both time scales, defining a "+b" (unvoiced) or "+s" (voiced) LM. Or it may fall, likewise: "-b" or "-s", respectively. In practice, it often happens that power rises in three or four frequency bands but stays nearly constant (to within 6 db) in the remaining ones.

A more complicated case arises for fricative-like "f" (unvoiced) or "v" (voiced) onset and offset LMs. Here, the power rises at high frequencies and simultaneously *falls* at lower frequencies, defining a "+f" or "+v". Or it may do the opposite, i.e., falling at high frequencies and rising at low ones: "-f" or "-v", respectively.

Note that "b"/"s" LMs always take precedence over "f"/"v". That is, if power rises in at least three bands, then SpeechMark detects a "+b"/"+s"; a "f"/"v" LM is *not* detected even if power falls in the other bands. And likewise for power falling in at least three bands: SpeechMark detects "-b"/"-s".

Figure 1 shows an example of the abrupt LMs for one syllable of an infant babble. In contrast to peak LMs, SpeechMark functions draw abrupt LMs above the waveform, labeled by lowercase letters. SpeechMark groups the LMs into one syllabic cluster, covering exactly the segment from the beginning at +g to the ending –g, but (in this example) not beyond. However, they are also grouped into an utterance that does extend beyond this point.

Also notice that the narrow-band spectrogram shows the characteristic horizontal stripes of well-developed voicing. However, it further shows two abrupt changes of period at 0.04s and 0.26s, as well as a loss of periodicity (-p) at 0.33s. Finally, notice that an acoustic event at 0.17s is (correctly) *not* detected as a LM, because it does not appear in enough spectral bands.

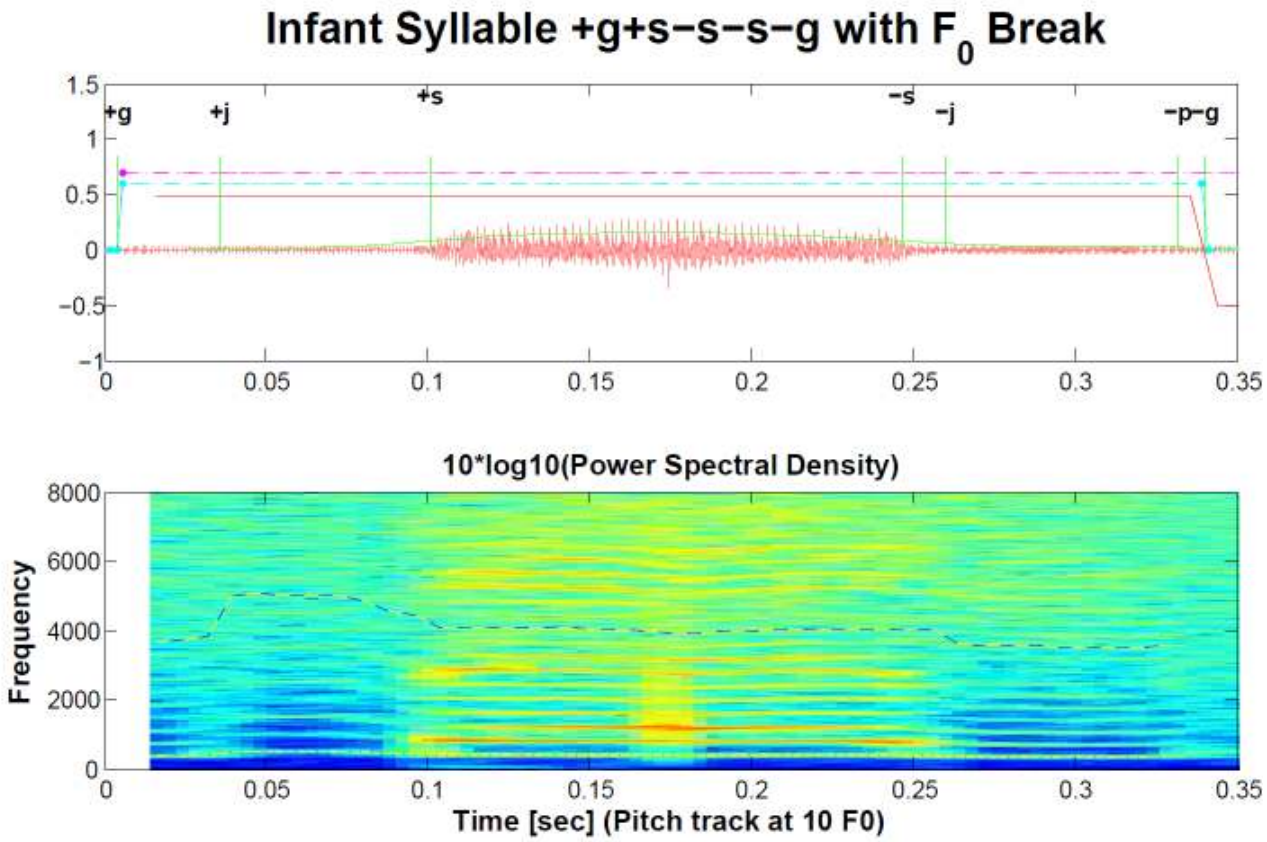

Figure 1. Example landmarks. One syllable of an infant babble is shown. The LMs are placed at instants of abrupt change of energy occurring simultaneously across multiple frequency ranges and at multiple time scales. (*top*) Waveform with smoothed amplitude envelope, landmarks (+g through –g, *green vertical lines*), and landmark grouping. Graphics show the interval of voicing (*solid red line*), grouping as a syllabic cluster (*dashed light blue*), and grouping as part of an utterance that continues beyond the window (*dashed magenta*). (*bottom*) Narrow-band spectrogram of the segment with dotted line through F_0 and dashed line through 10 F_0. The spectrogram shows the harmonics (horizontal stripes). Periodicity is strong even at the start of voicing (0.01s), so the +*g* LM is coincident with the corresponding +*p* (not shown). Note that the event at 0.17s affects too few spectral bands and therefore does not generate a LM. Also note abrupt jumps in F_0: Jumps and periodicity events do not contribute to defining the syllabic cluster, so this example is considered a +g+s-s-s-g cluster.

The following table summarizes the rules for the abrupt LMs.

Table. Rules to identify each type of AC LM. The symbols and mnemonics are *not* intended to identify underlying articulatory or phonetic events, only to suggest examples: syllabic, voiced frication, etc.

Symbol	Mnemonic	Rule
+g	Glottal onset	Beginning of sustained laryngeal vibration, i.e., of periodicity or of power and spectral slope similar to that of a nearby segment of sustained periodicity
-g	Glottal offset	End of sustained laryngeal motion
+p	Periodicity onset	Beginning of sustained periodicity of appropriate period
-p	Periodicity offset	End of sustained periodicity of appropriate period
+j	F_0 jump upward	Abrupt upward jump in F_0 by at least 0.1 octave (approx.)
-j	F_0 jump down	Abrupt downward jump in F_0 by at least 0.1 octave (approx.)
+b	Burst onset	At least 3 of 5 frequency bands show simultaneous power increases of at least 6 dB in both the finely smoothed and the coarsely smoothed contours, in an unvoiced segment (not between +g and the next -g)
-b	Burst offset	At least 3 of 5 frequency bands show simultaneous power *decreases* of at least 6 dB in both the finely smoothed and the coarsely smoothed contours, in an unvoiced segment
+s	Syllabic onset	At least 3 of 5 frequency bands show simultaneous power increases of at least 6 dB in both the finely smoothed and the coarsely smoothed contours, in a voiced segment (between +g and the next -g)
-s	Syllabic offset	At least 3 of 5 frequency bands show simultaneous power *decreases* of at least 6 dB in both the finely smoothed and the coarsely smoothed contours, in a voiced segment
+f	Frication onset	At least 3 of 5 frequency bands show simultaneous 6-dB power increases at high frequencies *and* decreases at low frequencies (unvoiced segment)
-f	Frication offset	At least 3 of 5 frequency bands show simultaneous 6-dB power *decreases* at high frequencies and increases at low frequencies (unvoiced segment)
+v	Voiced frication onset	At least 3 of 5 frequency bands show simultaneous 6-dB power increases at high frequencies *and* decreases at low frequencies (voiced segment)
-v	Voiced frication offset	At least 3 of 5 frequency bands show simultaneous 6-dB power *decreases* at high frequencies and increases at low frequencies (voiced segment)