Peak Landmarks in *SpeechMark*

J. MacAuslan, S. Boyce

Aug. 15, 2016

Landmarks (LMs) are acoustically identifiable points in an utterance. They come in the form of abrupt transitions (*abrupt* LMs) and peaks (*peak* LMs) of some contour or contours. Here we describe the peak set of landmarks used in SpeechMark®.

Vowel Peak Landmarks: Speech is characterized by sequences of energy rises and falls, corresponding roughly to syllabic units. The energy peaks are termed *syllabic nuclei*. These are typically monophthongal or diphthongal vowels, but may also be semivowels or syllabic consonants such as certain occurrences of /r/, /l/ or /n/ (as in *bird*, *bottle*, *button*). In what follows, we will use the term "vowel" or "vowel landmark" (V LM) to include the case of syllabic peaks derived from both vowels and syllabic consonants.

An important class of peak-type landmarks marks the nucleus of such syllabic units. These are termed *vowel peak landmarks* or *vowel landmarks.* The SpeechMark function to locate these is named *vowel_lms*, and *vowel_segs_full* or *vowel_segs_std* may be used to combine these vowel LMs with abrupt LMs to identify well-formed segments of speech. Vowel landmarks in SpeechMark are defined by a local peak of harmonic power.[1] Articulatorily, vowel landmarks often correspond to the maximum opening of the mouth within a syllabic unit.

Frication Peak Landmarks: Speech is also characterized by the onset and offset of air turbulence generated in the vocal tract. Sustained speech sounds with this characteristic are termed *fricatives* (or spirants) and the sound itself is termed *frication noise*. Articulatorily, frication noise is generated when air passes through a narrow opening in the vocal tract, a vocal tract constriction, which causes the airflow to become turbulent. The constriction may be caused by two articulators approaching one another (as with the lips or the vocal folds), or it may be caused by a single articulator approaching a stationary structure (as when some part of the tongue approaches the palate or teeth). Identifying such occurrences (F LMs) is the purpose of the SpeechMark *fricative_lms* function.[2]

Because the total energy passing through the vocal tract into the outside air is reduced when it encounters a narrow constriction, frication noise is necessarily characterized by weaker total energy than that of vowels. The energy is also more broadly distributed over frequency. At low frequencies, it is considerably weaker than vowel energy. Fricative noise may be strong at high frequencies. In practice, SpeechMark treats 0-500 .

Strong frication is generated as a consequence of well-developed air turbulence. Remarkably, the fractal dimension of *all* well-developed turbulence is 5/3 when measuring scalar signals such as pressure. This fact is the consequence of a universal property of turbulence, the Kolmogorov spectrum.

[1] Such landmarks are similar to those of Howitt, A.W., *Automatic Syllable Detection for Vowel Landmarks*, doctoral thesis M.I.T., Cambridge, MA. 2000.

[2] In principle, similar remarks apply as well to aspiration, sustained air turbulence generated by a constriction in the glottis rather than the oral cavity. However, it is unknown whether or under what conditions *fricative_lms* reliably detects such intervals.

A fractal dimension can be thought of as the degree of roughness or smoothness in the graph or plot of the waveform over time. The smoothest possible signal will have a fractal dimension of one, a smooth curve. The roughest possible signal will fill the entire two dimensional plot area and have a fractal dimension of two.

Perhaps surprisingly, computing fractal dimension is fast and simple. It depends solely on the difference between the maximum and minimum values of the signal over 2, 4, 8, and 16 samples (for instance).[3]

The SpeechMark function *audio_fractaldim16* computes a fractal dimension contour instant by instant over intervals of approximately one msec (specifically for 16-kHz signals). Therefore, the fractal dimension measures the dimensionality of the signal around 2 kHz (8 samples at 16 kHz), ensuring that multiple signal values are available for comparison at a timescale of approximately ½ msec.

The primary indicator of the peak of frication is a fractal dimension as close to 5/3 as possible. However, if a signal has a fractal dimension of 5/3 but has high power in low frequencies, we do not consider this a fricative. It may be air turbulence (perhaps environmental) but it cannot be assumed to be a component of speech. Fricative turbulence is produced by a narrow vocal-tract constriction, which automatically imposes a reduction in energy at low frequencies.

Brief and voiced frication both often have fractal dimensions somewhat less than 5/3, typically approximately 1.5. In contrast, vowels have fractal dimensions between 1 and 1.2 (1 being the lowest possible value for fractal dimensionality): e.g., that of the vowel in *fit* in Figure 1. Artificial signals including those of clipped audio can have fractal dimensions close to 2, the upper limit.

The figures show examples of the fractal-dimension and high-vs.-low frequency energy contours defining frication as well as the specific points – the landmarks -- that are identified as the local peaks of those contours. In Figure 1, notice that the stop release is so strong (perhaps hyper-articulated) that the resulting burst shows fully developed turbulence and an F landmark. This might seem surprising, because *fit* ends with /t/, a stop consonant, not a fricative. As is true throughout the SpeechMark suite, however, this tool analyzes the speech *as produced*, not necessarily as expected or even as intended.

Figure 1 illustrates one other point: These two contours may have meaningless values during near-silent intervals. Therefore, all processing suppresses detection during such intervals, to avoid finding F LMs before and after the production. (This is accomplished with a simple total-energy contour, not shown in the figures.)

[3] "Fractal dimensions of speech sounds", Maragos & Potamianos 1999, *J. Acoustical Soc. Of America* **105**(*3*), pp. 1925-1932.

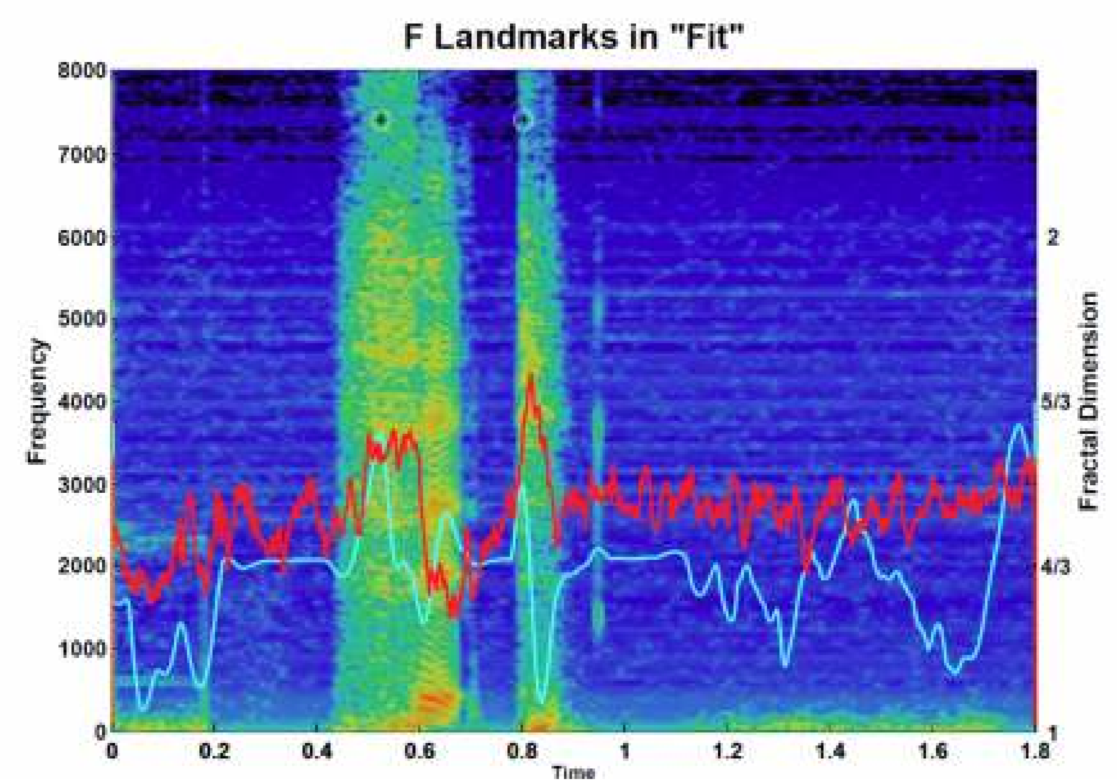

Figure 1. Spectrogram, Contours and Frication (F) Landmarks in "Fit". Note the vowel produced during 0.60-0.65s and the prominent stop release at 0.80-0.85s. (*red*, right scale) Contour of fractal dimension. (*light blue*, no scale) Contour of energy contrast, high-frequency minus low-frequency. (Scaling: 0 => LF dominant; 4000 => HF dominant.) *Not shown*: total-energy contour to suppress detection in near-silent intervals. (*white+black marks* shown at 7500 Hz, left scale) F landmarks detected at 0.52s, 0.80s.

In Figure 2, we see the importance of using both the dimension and contrast contours. For example, strong contrast peaks occur at 2.30s (*air*) and 2.75s (*act*), due to sonorants with high energy at high frequencies; however, the fractal dimension is appropriately very low there, so no landmarks are detected. Conversely, at 2.9s (*act*) and 3.1s (*like*), strong though brief bursts produce fractal-dimension peaks; but with only low contrast between high- and low-frequency energy, these do not produce landmarks.

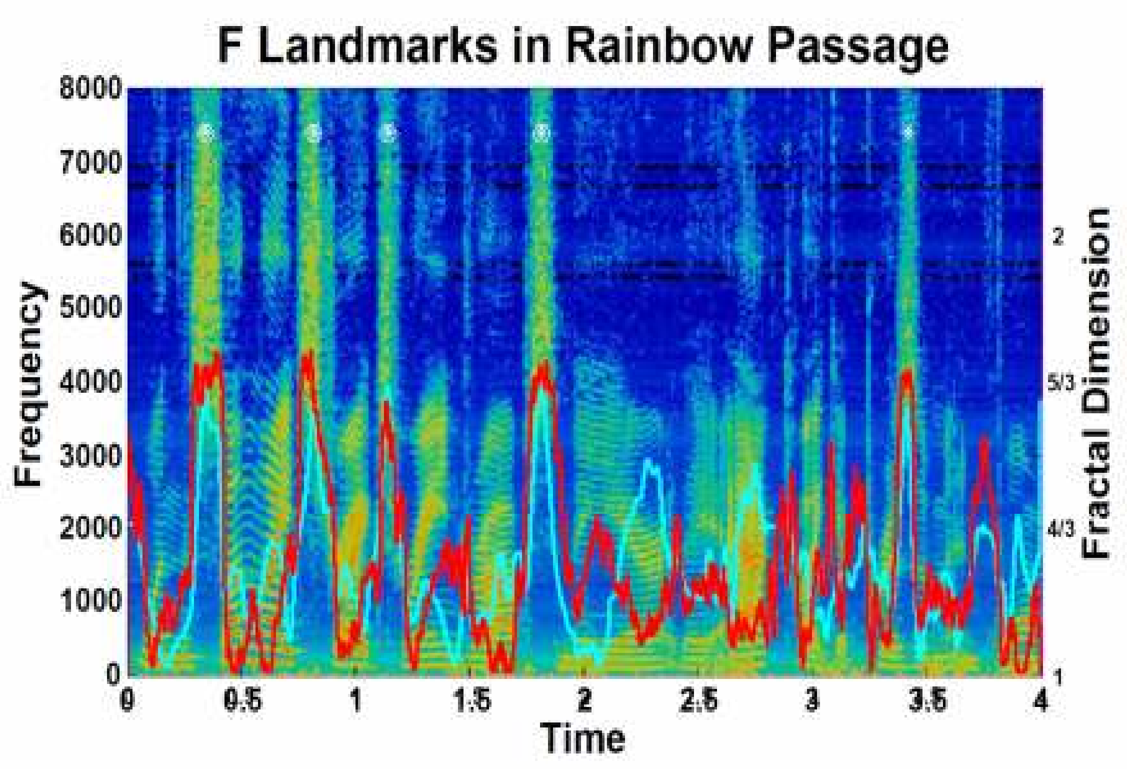

Figure 2. Spectrogram, Contours and Frication (F) Landmarks in Sentence. The sentence is "When the sunlight strikes raindrops in the air, they act like a prism and form." Contours and F landmarks (0.4, 0.8, 1.1, 1.7, 3.4s) as in Figure 1. Note that strident fricatives, both voiced (*priSm*, 3.4s) and unvoiced (*Sunlight*, 0.4s; *StrikeS*, 0.8 and 1.1s; *raindropS*, 1.7s), have sufficient airflow and duration for well-developed turbulence, whereas other points (*THe*, 0.3s; *Form*, 3.8s) and bursts (*aCT*, 2.9s; *liKe*, 3.1s) may not.