SpeechMark: Landmark Detection Tool for Speech Analysis

Suzanne Boyce[1]*, Harriet Fell*[2]*, Joel MacAuslan*[3]

[1]Department of Communication Sciences and Disorders, University of Cincinnati, Cincinnati, OH, USA
[2]College of Computer and Information Science, Northeastern University, Boston, MA, USA
[3]Speech Technology and Applied Research, 54 Middlesex Turnpike, Bedford, MA, USA

boycese@ucmail.uc.edu, fell@ccs.neu.edu, joelm@staranalyticalservices.com

Abstract

Landmark-based software tools are particularly suited to fast, automatic analysis of small, non-lexical differences in production of the *same speech material* by the *same speaker*. We are building a suite of independent applications and plugins as toolkits that make our landmark-based software system, SpeechMark, available to the wider scientific community. This will be achieved by extending existing software platforms with “plug-ins” that perform specific measures and report results to the user and by developing a MATLAB toolkit. These tools provide automatic summary statistics for measures of speech acoustics based on Stevens’ paradigm of landmarks, points in an utterance around which information about articulatory events can be extracted.

Index Terms: speech production, articulation, landmark, software.

1. Introduction

Changes in speech articulation can be used to detect, track and measure changes in motor coordination due to speaking style, health status, or mental operations such as working memory, motor planning, or integration of auditory and proprioceptive feedback. Further, speech recordings are (a) non-invasive, (b) inexpensive to collect, and (c) easily integrated into existing research and clinical protocols.

However, software for the scientific use of speech data is typically written for specific purposes and disseminated “as is” through informal open-source networks. Thus, progress has been handicapped by lack of access to software that (a) will produce automatic measures, and (b) is available in a robust, user-friendly form that can be adapted to different scientific aims and environments.

In previous work, we have developed an automatic system for detecting and measuring *landmarks*, i.e., acoustic events that correlate with changes in speech articulation [1]. Most research using the landmark approach has focused on the lexical content of speech [2, 3]. In our work, we have found that tools based on landmarks can be useful for investigating non-lexical attributes of speech, such as syllabic complexity or vowel space area over time. In particular, we have found that landmark-based software tools are particularly suited to analysis of small differences in production of the *same speech material* by the *same speaker*. This more-limited landmark approach has been useful for studies across a wide variety of disorders and behaviors: e.g. emotional change, the development of motor control in children’s speech, Parkinson’s Disease, and speech in response to sleep deprivation, among others [4-9]. Software enabling investigation of the landmark approach, however, has not been available to the research community at large.

As part of a three-year U.S. NIH-funded project, we are making our acoustic landmark detection tool, the SpeechMark system, available to the broader scientific community. SpeechMark will be available in two forms: (1) as a MATLAB toolbox for software developers, and (2) as a “plug-in” or “port” designed to augment the capabilities of software already in use by different sectors of the scientific community. We hope that availability of easy-to-use tools based on acoustic landmark detection will enable scientists to be more productive across a wide range of behavioral and biomedical research.

2. Landmarks Reflect Articulation

Landmark analysis is based on the fact that different sounds produce different patterns of abrupt changes in the acoustic signal simultaneously across wide frequency ranges.   For instance, the abrupt increase in amplitude for a broad range of frequencies above 3 kHz can be used to indicate the onset of bursts. Likewise, an abrupt decrease in the same frequency bands can be used to indicate the end of frication.   The use of onset and offset data in other frequency bands can be used to indicate sonorancy; i.e., intervals when the oral cavity is relatively unconstricted. Vowel landmarks represent local energy maxima characterized by harmonic power. These landmark patterns are identified by comparison between “coarse” and “fine” spectral detail.

This system makes no attempt to identify phonemes, but it is sensitive to broad categories of speech sounds and to aspects of metrical structure. An important aspect of the technique relies on setting empirically derived thresholds for the detection of abrupt acoustic changes in specified frequency bands. Recall that changes in the acoustic signal occur simultaneously across wide frequency ranges. When the onset of energy does not exceed threshold in a particular frequency band, i.e., is not quite abrupt enough to trigger the detection of a landmark, no landmark may be assigned. Thus, small acoustic differences in the way the same speech material is produced (i.e. in different styles or under different conditions) will reveal themselves as different patterns of landmarks.

Unlike systems focused on speech recognition, which involve detection of a large range of landmarks, our system focuses especially on detecting abrupt landmarks [10] and vowel landmarks [11]: **g(lottis)** - onset (+g) or offset (-g) of voicing; **s(yllabicity)** - onset (+s) or offset (-s) of voiced sonorant consonants; **b(urst)** - onset (+b) of the burst of air following stop, affricate consonant release, or onset of frication noise for fricative consonants and offset (-b) - where aspiration or frication noise ends abruptly due to a stop closure; **V(owel)** - peak harmonic power in a sonorant region**.** The system also notes energy changes associated with patterns of frication, denoted by +/-v and +/-f.

3. Syllabic Cluster Analysis

To date, we have applied the SpeechMark system primarily to detect changes in articulatory precision as a result of speaking style [12], disease state [6], and sleep state [14], or to detect syllabic complexity as a measure of articulatory coherence in development [13]. In this paper, we describe two studies using our syllabic complexity measure. In the one case, we applied this measure, termed the Syllabic Cluster Analysis, to speech produced by Parkinson's Disease patients undergoing Deep Brain Stimulation (DBS) therapy [6]. In the second case, we applied this measure to speech produced by healthy speakers in rested and sleep-deprived conditions.

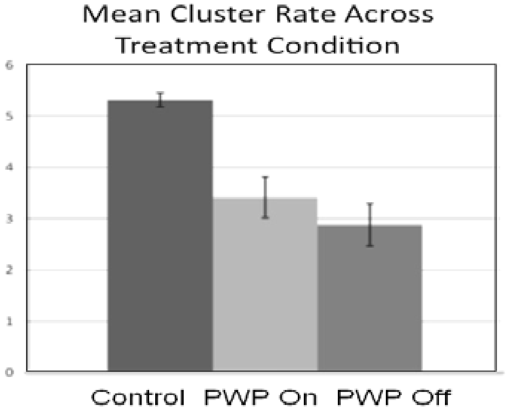


Figure 1. The mean syllabic cluster rate for 12 speakers with Parkinson's Disease (PWP) vs. 15 matched control subjects (Control) in two conditions: (1) Stimulator ON, and (2) Stimulator OFF.

It should be noted that the sense of the term "syllabic complexity", as we use it here, refers to the speech signal as uttered. A word such as "interesting", for instance, can have four syllables in its canonical form, but when uttered as /ɪntrɛstɪŋ/ it can be said to have three syllables. If uttered in a canonical fashion, both pronunciations will show a characteristic pattern of landmarks for each syllable, and as long as the syllables are uttered with the same acoustical characteristics the same pattern of landmarks will be detected. However, if the syllables are uttered less canonically—perhaps with less extreme articulatory movements, less precise timing, or reduced aerodynamic support, fewer landmarks will be detected. The measure thus reflects two effects: (1) fewer syllables and (2) simplification of multi-element constituents such as syllable onsets and rimes.

The Syllabic Cluster analysis works by grouping sequences of detected landmarks into clusters that correspond, roughly, to syllabic units in the acoustic speech signal. The grouping rules include categorical dependencies as well as dependencies of timing, and were empirically determined from datasets of speech. For example, a gap of 30 ms in voicing, with whatever ±b's immediately follow it, is one type of syllable cluster endpoint. Landmarks that do not conform to syllable cluster rules are typically suppressed as non-speech noise. For example, burst-like noise that does not occur within 120 ms before a voiced region, or 80 ms after, is not included as part of a cluster and will be suppressed. The syllabic grouping procedures are described in more detail in [14].
Some examples of the more common types of syllabic cluster are:

- (+g,-g)- singleton V or CV syllables, where C is voiced.
- (+b,+g,-g) – CV syllable beginning with fricative: (+b) marks the presence of frication.
- (+b,-b,+g,-g) - syllables with an initial plosive: (+b, -b) mark the beginning and end of the release.

The usefulness of the Syllable Cluster Analysis measure in SpeechMark has been tested in a number of studies. Below we compare speech produced by the same speakers in two different areas of scientific inquiry: (a) Parkinson's Disease patients who were receiving Deep Brain Stimulation (DBS) and healthy speakers in a rested vs. sleep-deprived condition.

In the typical progression of Parkinson's Disease, patients show clinically significant levels of unintelligible speech later than they show gross motor symptoms. Thus, patients in DBS programs may not be showing clinically overt signs of dysarthric speech. However, the application of DBS therapy can sometimes cause their speech intelligibility to worsen in subtle ways. Sleep deprivation is commonly thought to cause "slurring" of speech but the effects are also subtle [15].

For the study shown in Figure 1, subjects produced a sequence of multiple rapid repetitions of the syllable "ka" with the DB stimulator "ON" vs. "OFF". Because there were inconsistent numbers of repetitions in the two conditions, the data are reported in terms of mean cluster

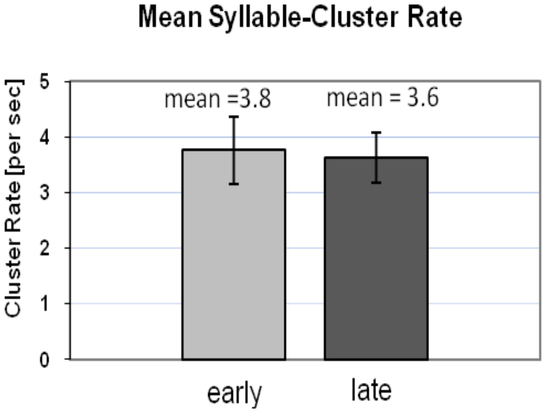


Figure 2. The mean rate of Syllabic Cluster occurrence for 17 speakers of American English reading the Rainbow Passage aloud in Early vs. Late sessions of a 30-40 hour period without sleep.

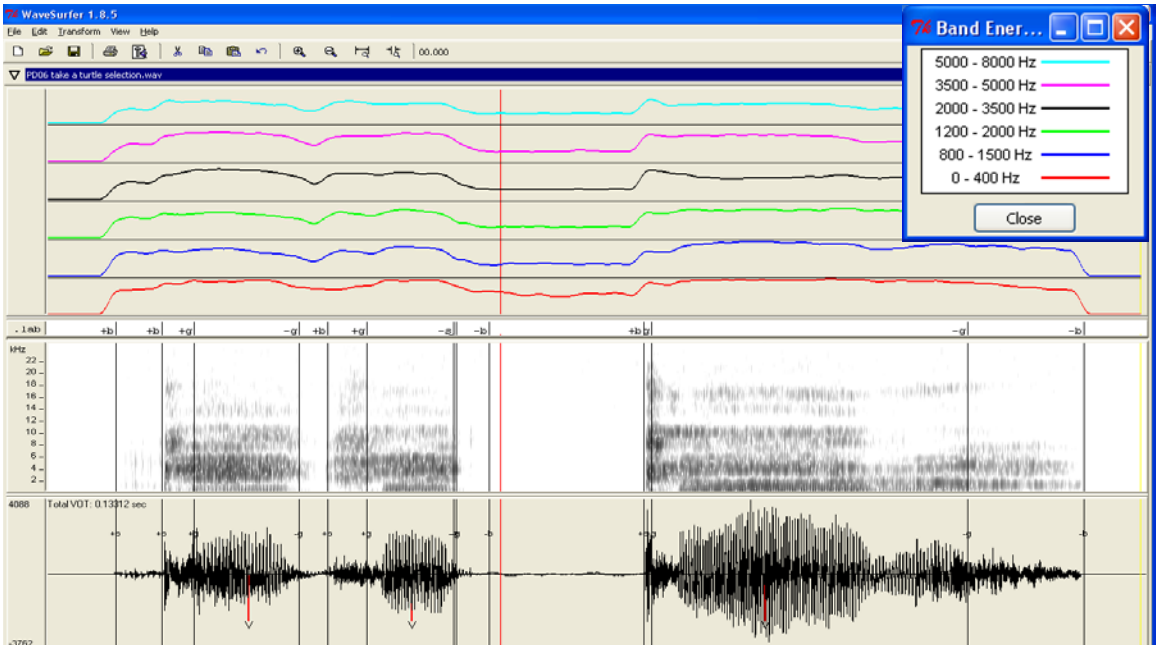
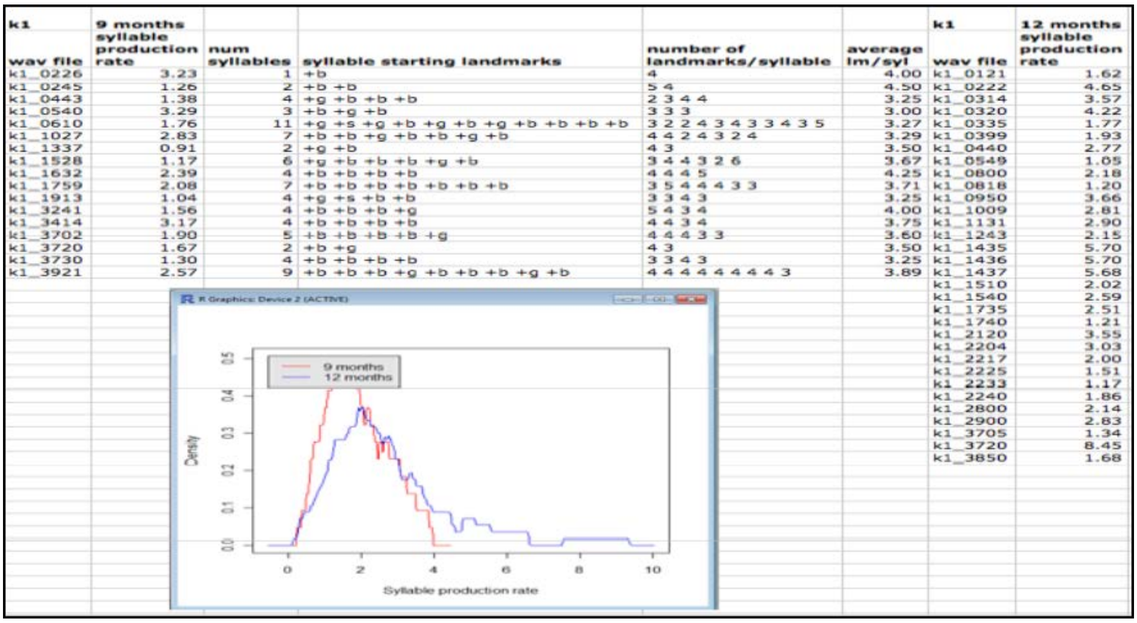


Figure 3*: Wavesurfer display showing an audio recording of "Take a turtle" by a speaker of American English with Parkinson's Disease. Landmarks are shown on the waveform pane and also in a transcription pane above the spectrogram pane. The orthographic transcription is shown on top.*



k1	9 months					k1	12 months
wav file	syllable production rate	num syllables	syllable starting landmarks	number of landmarks/syllable	average lm/syl	wav file	syllable production rate
k1_0226	3.23	1	+b	4	4.00	k1_0121	1.62
k1_0245	1.26	2	+b +b	5 4	4.50	k1_0222	4.65
k1_0443	1.38	4	+g +b +b +b	2 3 4 4	3.25	k1_0314	3.57
k1_0540	3.29	3	+b +g +b	3 3 3	3.00	k1_0320	4.22
k1_0610	1.76	11	+g +s +g +b +g +b +g +b +b +b +b	3 2 2 4 3 4 3 3 4 3 5	3.27	k1_0335	1.77
k1_1027	2.83	7	+b +b +g +b +b +g +b	4 4 2 4 3 2 4	3.29	k1_0399	1.93
k1_1337	0.91	2	+g +b	4 3	3.50	k1_0440	2.77
k1_1528	1.17	6	+g +b +b +b +g +b	3 4 4 3 2 6	3.67	k1_0549	1.05
k1_1632	2.39	4	+b +b +b +b	4 4 4 5	4.25	k1_0800	2.18
k1_1759	2.08	7	+b +b +b +b +b +b +b	3 5 4 4 4 3 3	3.71	k1_0818	1.20
k1_1913	1.04	4	+g +s +b +b	3 3 4 3	3.25	k1_0950	3.66
k1_3241	1.56	4	+b +b +b +g	5 4 3 4	4.00	k1_1009	2.81
k1_3414	3.17	4	+b +b +b +b	4 4 3 4	3.75	k1_1131	2.90
k1_3702	1.90	5	+b +b +b +b +g	4 4 4 3 3	3.60	k1_1243	2.15
k1_3720	1.67	2	+b +g	4 3	3.50	k1_1435	5.70
k1_3730	1.30	4	+b +b +b +b	3 3 4 3	3.25	k1_1436	5.70
k1_3921	2.57	9	+b +b +b +g +b +b +b +g +b	4 4 4 4 4 4 4 4 3	3.89	k1_1437	5.68
						k1_1510	2.02
						k1_1540	2.59
						k1_1735	2.51
						k1_1740	1.21
						k1_2120	3.55
						k1_2204	3.03
						k1_2217	2.00
						k1_2225	1.51
						k1_2233	1.17
						k1_2240	1.86
						k1_2800	2.14
						k1_2900	2.83
						k1_3705	1.34
						k1_3720	8.45
						k1_3850	1.68

Figure 4: *Distribution of landmark information for different files or folders, as produced by our prototype R package.*

count over time. As Figure 1 shows, Parkinson's patients produced fewer syllable clusters in the Stimulator ON condition. The differences were significant at the .01 level. This result matches clinical impressions of reduced intelligibility in the Stimulator ON condition for these patients.

In the Sleep Deprivation study, the speech of 17 speakers of American English (9 female, 8 male) was recorded at 8 hour intervals over 30-35 hours without sleep. Subjects read aloud the Rainbow Passage each time. As Figure 2 shows, there was a significant difference in syllabic cluster rate between the first and last sessions (p<.05, binomial test), with later sessions showing fewer clusters. As with the Parkinson's Disease study, these results parallel results from a study of clear vs. conversational speaking style [12], where the same speakers using the same speech materials produced fewer syllable clusters in the less intelligible speaking style.

4.1. Development of MATLAB Code

The core SpeechMark computational engine is already implemented in MATLAB. Because the MATLAB platform itself is costly, especially for non-academic users, we are developing a suite of independent applications and plugins as toolkits that run within existing software packages. Plugins for the WaveSurfer and R open-source packages have already been developed and are undergoing beta-testing. A laboratory-grade version of the MATLAB toolbox is currently being documented and should be available via a website for beta-testing in fall 2012. This will add user controls for thresholding landmark detection in the presence of noise and for changing the frequency bands used in the SpeechMark analysis. Plugins for Microsoft EXCEL and PRAAT are in the planning stage. We anticipate that these toolkits will be available gratis or at a modest price.

4.2. Development of Plugin for Wavesurfer

We have produced a beta-test version of our landmark detection system as a plugin to the Wavesurfer speech analysis platform. This plugin is designed for researchers with a primary interest in analyzing the placement of landmarks of each type, patterns of clustering, or identification of non-speech sounds to be excised (See Figure 3). This version has user controls ("widgets") to produce automated measures or types of analyses for speech re-search such as:

- Scatterplot of F1 vs F2 at each +V landmark
- Detection of non-harmonic (and harmonic) voicing.
- Identification and suppression or removal of stray sounds, i.e., non-speech.
- Grouping of landmarks into syllable-like clusters.
- Time from +b to +g landmarks (similar to VOT).

And for instructional purposes or further development:

- Band Energies used in the landmark computation.

4.3. Development of Plugin for R

R is a powerful open-source statistical software system. We have implemented a beta-test version of a package for R that allows users to acquire landmark-based measures from multiple files and from multiple directories. These measures are then deposited into files according to the original directory structure design. The R package exports the list of landmarks for each file to the R environment for further statistical analysis. We have tested this feature of the package on a dataset containing e.g. many speech recording files from a set of different infants, organized in multiple directories, each containing several subdirectories of recordings for the same infants at different ages.

Figure 4 shows an example of different distributions of landmark information for different files or folders, as produced by our prototype R package.

4. Future Plans

As noted above, plans are underway to develop SpeechMark plugins for Microsoft Excel, and possibly PRAAT, in future. In addition, we are conducting formal studies with beta-testers of the software to evaluate the usability of our tools, and soliciting input from user communities about the features they would like to see in these tools. We invite members of the scientific community who are interested in evaluating our beta software to contact us.

5. Conclusion

One barrier to greater use of speech analysis in scientific investigation is the lack of user-friendly automatic measurement methods. We have developed a set of software tools based on objective detection and classification of acoustic landmarks and clusters that produces statistically reproducible analyses of speech characteristics in real time. We expect to make these tools available in different forms for use by the scientific community.

6. Acknowledgements

This work was funded by grants R43 and R44 DC010104, R42 AG033523, R21 HL086689, and R42 HD34686 from the United States National Institutes of Health (NIH).

7. References

[1] Stevens, K.N., "Toward a model for lexical access based on acoustic landmarks and distinctive features". Journal of the Acoustical Society of America, **111**(4): 1872-91, 2002.

[2] Juneja, A. and C.Y. Espy-Wilson. "Speech Segmentation Using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines", in International Joint Conference on Neural Networks Proc., 2003.

[3] Slifka, J.S., et al. "A Landmark-Based Model of Speech Perception: History and Recent Developments", in From Sound to Sense: Fifty Years of Speech Research, 2004.

[4] Fell, H., et al., "Vocalization Age as a Clinical Tool", in International Conference on Spoken Language Processing (ICSLP) Proc., 2002.

[5] Fell, H.J., et al., "Using early vocalization analylsis for visual feedback", 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA) Proc., 39-42, 2003.

[6] Chenausky, K., J. MacAuslan, and R. Goldhor, "Acoustic Analysis of PD Speech". Parkinson's Disease, 1-13, 2011.

[7] Boyce, S., A. Bradlow, and J. MacAuslan, "Landmark analysis of clear and conversational speaking styles", in 150th Meeting of the Acoustical Society of America, 2005.

[8] Boyce, S., et al., *Landmark-based Analysis of Sleep Deprived Speech.* Acoustics '08 Proc., 2008.

[9] Boyce, S., et al., "Automated Tools for Identifying Syllabic Landmark Clusters that Reflect Changes in Articulation", in 7th Annual Workshop for Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA) Proc., 2011.

[10] Liu, S., "Landmark detection in distinctive feature-based speech recognition", Ph.D. Dissertation, Massachusetts Institute of Technology: Cambridge, Massachusetts, 1995.

[11] Howitt, A.W., "Automatic Syllable Detection for Vowel Landmarks", Ph.D. Dissertation, Massachusetts Institute of Technology: Cambridge, Massachusetts, 2000.

[12] Boyce, S., et al., "Automatic Detection of Differences Between Clear & Conversational Speech" in American Speech-Language-Hearing Convention, 2007: Boston, MA.

[13] Boyce, S., W. Carr, and J. MacAuslan, "Effects of Sleep Deprivation on Speech Articulation and Intelligibility in Noise" in Aerospace Medical Association (ASMA) Proc., 2011.

[14] Fell, H.J. and J. MacAuslan, *Vocalization Analysis Tools.* 4th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA Proc. 39-42, 2005.

[15] Harrison, Y. and J. Horne, "Sleep Deprivation Affects Speech." Sleep: Journal of Sleep Research & Sleep Medicine, Vol 20(10), 871-877, 1997.