Landmark-based Analysis of Sleep-Deprived Speech

Suzanne Boyce*, Joel MacAuslan+, Ann Bradlow#, Rajka Smiljanic#

*University of Cincinnati, +Speech Technology and Applied Research Corp., #Northwestern University

5aScc31

Introduction

There is a common perception that speech articulation becomes "slurred", or less precisely articulated, under sleep deprivation conditions. There have been few studies of speech under sleep deprivation. Morris et al. (1960) and Harrison & Horne (1997) found that listeners heard a difference between speech recorded under rested and sleep-deprived conditions. However, their measures bear only an indirect relation to articulatory clarity per se.
Speech researchers have identified a number of measures that distinguish clearly articulated speech from less-clearly articulated speech (Bradlow et al., 2005, Krause & Braida, 2004, Chin & Pisoni,1997, among others).

EXPECTED CHARACTERISTICS OF REDUCED ARTICULATORY CLARITY

Reduced pitch range (i.e. speech is more monotonic)
Reduced vowel space (i.e. vowels less distinct from one another, more like "uh")
Voiceless stops sound more like voiced stops (e.g. "t" sounds more like "d", "k" more like "g")
Less precise fricatives (e.g. "s" sounds more like "sh")
Unstressed syllables reduced or "swallowed", e.g. "plees" for "police", "inristin" for "interesting"

In past work, we have described the use of a "landmark"-based computer program to detect contrasts in articulatory clarity between "Clear" and "Conversational" speaking styles. In this paper, we test the hypothesis that rested and sleep-deprived speech will show changes in articulatory clarity similar to that seen in "Clear" vs. "Conversational" speech.

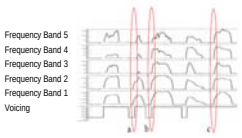


Figure 1: Illustration of Landmark Identification as patterns of abrupt changes in spectral bands. (*a*) Too few bands show large, simultaneous changes in energy. (*b*) All bands show large, simultaneous energy increases immediately before the onset of voicing, identifying a +b (burst) landmark. (*c*) All bands show large, simultaneous energy increases during ongoing voicing, identifying a +s (syllabic) landmark.

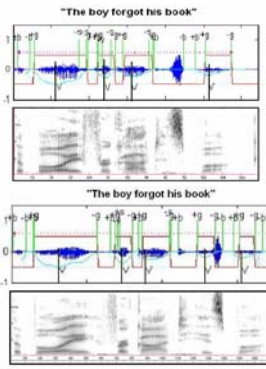


Figure 2. Examples of CONVERSATIONAL (top) and CLEAR (bottom) style productions of the same sentence by the same speaker. The top panel for each sentence shows the speech waveform, the bottom panel shows the spectrogram. Vertical lines indicate the points at which landmarks and vowel centers have been identified. Note that the sentence produced in CLEAR style production is longer. From Bradlow & Bent (2002) database.

Landmark Detection System (Fell & MacAuslan, 2003)

We use a form of the landmark analysis system of Liu (1995) based on Stevens (1991) that detects three types of landmarks:

1. q: glottis. Marks the time when the vocal folds transition from not vibrating to freely vibrating (+g) or vice-versa (-g). (Indicated from voicing band, seen at bottom of Fig. 1.)
2. s: syllabicity. Marks sonorant consonantal releases (+s) and closures (-s). These are always voiced.

3. b: burst. Designates frication onsets or affricate or stop bursts (+b) and points where aspiration or frication ends (-b) due to a stop closure. (Indicated from simultaneous abrupt changes in frequency bands.) These are never voiced.

The speech signal is automatically partitioned into 5 frequency bands plus voicing. Landmarks are identified as points where abrupt changes in the spectrum at particular frequency bands of a particular type coincide. As noted above, sequences of landmarks that represent syllabic groupings are then identified and tabulated.

NOTE: Our landmark system uses a threshold to determine if a landmark occurred. Thus, there may be evidence in the speech signal of a particular articulatory event, but if the evidence does not hit a threshold, the landmark will not be detected. Information regarding the 'strength' of a landmark is retained.

Landmarks can be used to eliminate pauses and to calculate most standard speech measures, such as VOT.

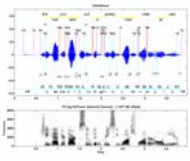


Fig. 3. Comparison of phonetically important landmarks marked by hand in the TIMIT database, and landmarks as detected by our system. The sentence is "She had your dark suit in greasy wash water all year". TIMIT transcription is at bottom.

Example Measures based on Landmark Analysis:

- Voice Onset Time (VOT). The total of time intervals between +b and +g landmarks. This is a measure of the difference between clearly articulated /t/, /k/ /p/ and /d/ /g/ /b/. Not robust to background noise.
- Total Number of Landmarks. Not robust to background noise.
- Syllabic Complexity. Landmarks can be grouped into clusters corresponding to syllabic units (a measure of syllabic complexity). Note that a CVC such as "cab" or "pat" may show up as (*+b-b+g-g+b-b*) when the consonants are clearly articulated, but as (*+b-b+g-g*) when the final consonant is unreleased or weakened. A syllable consisting of a single vowel (as in "a pout") will probably show up as (*+g-g*). This measure is very robust to background noise.
- Duration of "Strong" vs. "Weak" Syllables. The duration of "strong" vs. "weak" syllables is a rough measure of the degree to which syllables become reduced, or "swallowed".

Most Common Syllable Types [per PC Fit]

Rank	Label	Mean # per RP	Fraction	Mean Duration [ms]	Weight (Coeff)
1	+g-g	47.6	0.126	131	-1.6383
2	+g+s+s-g	31.0	0.082	61	-2.8756
3	+g-s-g	18.0	0.048	244	3.885
4	+g+s-g	17.8	0.047	200	-1.4729
5	+s-g	17.1	0.045	189	1.9392
6	+g+s	16.1	0.043	218	0.7519
7	+g+s-s-g	13.9	0.037	251	1.2793
8	+s-s-g	13.4	0.035	216	1.0626
9	+g-g+b	12.8	0.034	157	0.2932
10	+g-g-b	11.5	0.030	166	2.6692
11	+g-s	11.3	0.030	151	2.6641
12	+b+g-g	10.6	0.028	187	-3.6644
	Total of These	221.2304	0.586116		
	Total of ALL:	377.4516	1		

Table 1. Syllable Cluster Types ranked by frequency over total dataset, with weights according to the least-squares fit. The most common types are: +g+s-s, +b+g-s-g, +g+s-s-g, +g-s-g-b, +s-s-g, +b-b+g+s-g. A plus coefficient designates a syllabic cluster that is more common in Sleep Deprivation, a negative coefficient more common in the Rested case.

Three Most Common Syllable Cluster Types

Early
+b+g-g
+s-g-b
+g+s+s-g

Sleep-Deprived
+g+s-g-b
+g-s-g
+g+s-g+b

Clear
+g+s-s
+b+g-s-g
+g+s-s-g

Conversational
+g-s-g-b
+s-s-g
+b-b+g+s-g

Databases Analyzed

Clear vs. Conversational

(1) **Bradlow & Bent (2002)**
Speakers: 2 American English
Materials: 4 lists of BKB* sentences, 16 sentences each, plus additional list for female speaker
Listeners: 10 American English. Both subjects were less intelligible in CONVERSATIONAL than in CLEAR condition.

(2) **Boyce et al. (2007)**
Speakers: 10 American English
Materials: 6 lists of BKB sentences
Listeners: NONE

(3) **Smiljanic & Bradlow (2005)**
Speakers: 6 American English
Materials: 20 12-syllable sentences
Listeners: 30 American English

Sleep-Deprived vs. Rested

(1) WRAIR/NIDCD database (Carr unpublished*)
Speakers: 15 American English
Materials: Elicited Rainbow Passage (repeated at 8 hour intervals
Recording: Some background noise
Comparison: Early (12-14 hrs since sleep) vs. Late (36-42 hrs since sleep)

DCIEM database (Linguistic Data Consortium)
Speakers: 6 Canadian English (male)
Materials: Structured Conversation (Map Task) repeated in barrier format with different maps, listeners
Recording: Good
Comparison: Early (10 hrs since sleep) vs. Late (54 hrs since sleep)

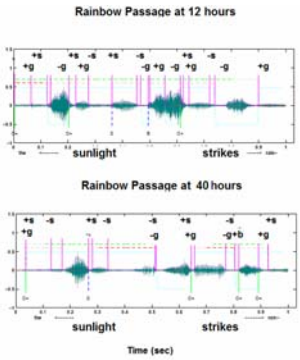


Fig. 4. The phrase "Sunlight strikes raindrops" from EARLY (~12 hours since last sleep) and LATE (~40 hours since last sleep) sessions. From NIDCD/WRAIR database.

Notable Aspects of Landmark Analysis

•Ignores formant frequencies except as total amplitude per (5) frequency bands
•Uses durational relationships between abrupt changes to determine landmarks
•Ignores nominal linguistic component (syllable, word, phrase, sentence)
BUT Landmark patterns reflect syllables AS UTTERED

Examples:
+g,-g ----- "aah", uttered in isolation.
+b,+g,-g -----"see"
+b,-b,+g,-g ----- crisply articulated "tea"
+b,+g,-g ----- less crisply articulated "tea"

Note: We assume that a stretch of no speech (i.e. no voicing) of 350 ms or more = a pause. Our measures are calculated AFTER pauses have been subtracted out.

Statistical Procedure for Syllable Cluster Principal Components Fit:

Lists of syllabic Clusters and their distributions were transformed to approximately Gaussian-distributed variables of zero mean and unit standard deviation (representing the irreducible noise level).
These lists were then analyzed as a batch to determine a least-squares fit to the data (a calibration of the coefficients). At this point, we suppressed all principal components contributing less than ~3% of the variance.

Point scores were derived from the goodness-of-fit (FIT) between the major principal components and the condition (Clear vs. Conversational, Rested vs. Sleep-Deprived) that the segment was drawn from. These point scores form an "estimated" or "constructed" variable representing the degree to which the data pattern around syllabic cluster distribution, and thus articulatory clarity.

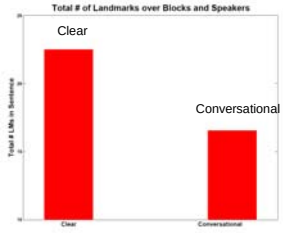


Figure 5a and 5b. Number of Landmarks averaged over speakers and blocks (for each list, all sentence #1's were averaged, all sentence #2's were averaged, etc.). Note the lack of overlap between categories (p < .001). Data from Boyce et al. (2007).

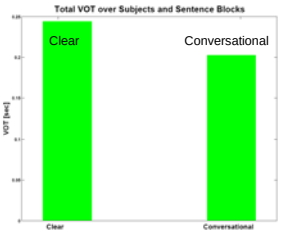


Fig. 6a and 6b. Total Voice Onset Time (VOT) averaged over Subjects and Sentence Blocks. Each line in Fig. 6b (below) shows mean over Blocks for each subject. From Smiljanic & Bradlow (2005) database.

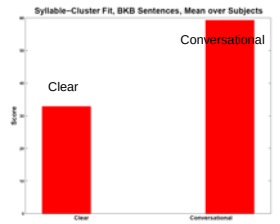


Fig. 7a and 7b. Syllabic Cluster Distribution for Boyce et al. 2007 Clear vs. Conversational Conditions, averaged over blocks of sentences (p < .001). Scores along estimated dimension reflect degree to which Principal Components Analysis shows relative frequencies of syllabic cluster types. Fig. 7b (below) shows means per block.

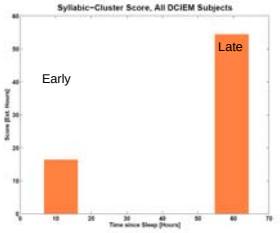


Fig. 8a and 8b-c (small). Syllable Cluster Distribution for DCIEM database. Each line is the mean per subject (p < .001). Fig. 8b and c (below) show individual subject data. From DCIEM database.

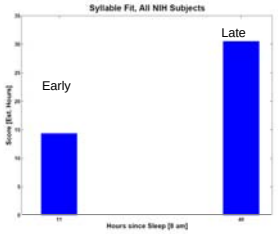
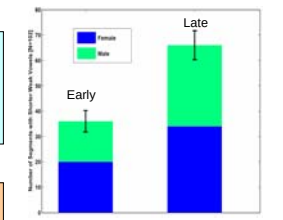
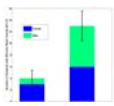


Fig. 9a and 9b (small). Syllable Cluster distribution for NIDCD/ WRAIR database (p < .001). In Fig. 9b, each line is the average per subjects.



Figs. 10a and 10b. Fig. 10a shows the number of 20-second segments of speech for which the mean duration of weak vowels is less than the mean duration of strong ones. "Strong" and "weak" are defined acoustically by same criteria in both conditions (p < .001). Fig. 10b (below) shows the # of subjects for whom the same is true.



Conclusion:

We conclude that sleep deprivation affects speech articulation, in a way parallel to the effect of sleep deprivation on the PVT task. The speech effects resemble those seen in other research on speech intelligibility (Bradlow & Bent, 2002; Krause & Braida, 2004) and are consistent with those reported in Harrison & Horne (1997) and Greeley (2007). These effects are very subtle and would not be noticeable to many listeners, but they appear to be both reliable and (automatically) measurable.

References:

Boyce, S., Bradlow, A., & MacAuslan, J. (2005). Landmark analysis of clear and conversational speaking styles. Paper presented at the150th meeting of the Acoustical Society of America, Minneapolis, Minnesota.

Bradlow, A. R., and Bent, T. (2002). The clear speech effect for non-native listeners. Journal of the Acoustical Society of America. 112 (1), 272-284.

Harrison, J. & Horne, Y. (1997). Sleep affects speech. Sleep Research 26:615.

Greeley, H.P. and Nesthus, T.E. (2007). Predicting Fatigue Using Voice Analysis. Aviation, Space, and Environmental Medicine, 76(7), 730-743.

Krause, J. C. & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. Journal of the Acoustical Society of America, 115, 362-378.

Liu, S. (1995). Landmark detection in distinctive feature-based speech recognition. Unpublished Ph.D. dissertation. Cambridge, MA.

Shriberg, L. & Kent, R. (1982). Clinical Phonetics. Allyn & Bacon:Boston.

Stevens, K .N. (1991). Speech perception based on acoustic landmarks: Implications for speech production. Perilus XIV. Papers from the symposium, Current phonetic research paradigms: Implications for speech motor control.

Fell, H. J., MacAuslan, J., Cress, C. J.,& Ferrier, L. J. (2003). Using early vocalization analysis for visual feedback. Proceedings of the 3rdth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA, 2003), 39-42.

Morris et al. (1960). Morris et al (1960) Arch. Gen. Psychiat. 2:247-254.

Smiljanic, R. & Bradlow, A. (2005). Production and perception of clear speech in Croatian and English. Journal of the Acoustical Society of America, 118 (3), 1677-1688.

Direct research support by NIH grant R21HL086689 "Measures of Sleep Deprivation from Speech Articulation" (to Suzanne Boyce), from U.S A.F. Contract F33615-02-M-6057, and from Walter Reed Army Institute of Research (WRAIR) and the National Institute of Deafness and Other Communication Disorders (NIDCD, internal division) is gratefully acknowledged. Much of the Landmark Detection technology was developed under NIH grants R42 DC00534 and R41 NS047959

We cite (with permission) analyses of unpublished data collected by colleagues Walter Carr, Ph.D., Tracy Rupp, Ph.D., Thomas Balkin, Ph.D. and Allen Braun, Ph.D. in a collaboration with the U.S. Navy, NIDCD and WRAIR. They are not responsible for any mistakes.