Using landmark detection to measure effective clear speech

Suzanne E. Boyce[1], Jean Krause[2], Sarah Hamilton[1], Rajka Smiljanic[3], Ann Bradlow, Ahmed Rivera Campos[1], Joel MacAuslan[4]
[1]University of Cincinnati, [2]University of South Florida, [3]University of Texas, [4]Northwestern University, [4]S.T.A.R. Corporation

Introduction

A number of studies have established that normal native speakers of a language know how to improve their intelligibility to listeners under intelligibility-challenging conditions. (Uchanski, 2005). This "Clear Speech" speaking style is significantly more intelligible to listeners; the average Clear Speech benefit is 15-17% to normal-hearing listeners in noise and to hearing impaired listeners in quiet (Uchanski, 2005). This is roughly the equivalent of a 5 dB improvement in signal/noise ratio (Uchanski, 2005). Many of these studies have reported that measures associated with articulatory "precision" are greater in Clear Speech, including: (1) an increase in vowel space, (2) increased consonant-to-vowel intensity differences, (3) stronger stop bursts, among others (Bradlow & Bent, 2002; Krause & Braida, 2002, 2004; Smiljanic & Bradlow, 2005).

A consistent finding in studies of Clear Speaking style is that there are significant speaker-to-speaker differences (Ferguson, 2004) and that some speakers are more intelligible than others when producing Clear Speech (Bradlow & Bent, 2002; Krause & Braida, 2004; Smiljanic & Bradlow, 2005). It is not immediately clear what distinguishes the better speakers, but one strong possibility is that these speakers produce more of the acoustic characteristics that distinguish Clear from Conversational Speech. The ability to detect when a speaker's speech patterns are mostly likely to be intelligible would obviously be helpful in training clinicians, teachers and public safety workers to be more effective communicators.

In previous work, we have explored the the use of the ***SpeechMark***^{TM} landmark-based computer program to detect the acoustic characteristics of Clear speaking style. Landmark-based speech analysis takes advantage of the fact that important articulatory events, such as the onset and offset of frication, voicing, etc. show characteristic patterns of abrupt change in the speech signal (Stevens et al., 1992). These patterns are detected by an automated computer system set to a specific threshold of change over time, and assigned to a particular type of landmark. The onsets and offsets of landmarks also allow for automatic detection of pauses, speaking time, and voicing time. We use a form of the landmark analysis system of Liu (1995) based on Stevens et al. (1992) that detects three types of abrupt landmarks plus landmarks corresponding roughly to the acoustic center of a vowel. The speech signal is automatically partitioned into 5 frequency bands plus a voicing band. Abrupt landmarks are identified as points where abrupt changes in the amplitude of particular frequency bands coincide in a specified pattern. These landmark patterns are identified by comparison between "coarse" and "fine" spectral detail.

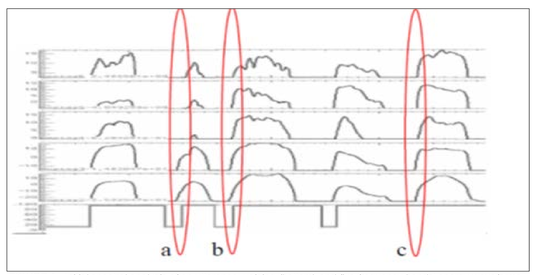

FIGURE 1. *Initial spectral analysis of an utterance: voicing (bottom) and five frequency bands' energy waveforms. (a) Too few bands show large, simultaneous changes in energy. (b) All bands show large, simultaneous energy increases immediately before the onset of voicing, identifying a +b (burst) landmark. (c) All bands show large, simultaneous energy increases during ongoing voicing, identifying a +s (syllabic) landmark.*

Hypotheses

In this paper, we focus on the problem of detecting the best (the most intelligible) exemplars of Clear Speaking style.

Phase I Hypothesis: Clear and Conversational speech recordings known to differ in intelligibility are characterized by different patterns of landmark-based measures.

Phase II Hypothesis: The BEST Clear Speech speakers will show a more extreme version of the landmark patterns that differentiate Clear from Conversational Speech.

Data set

Four collections of speech recordings and associated intelligibility data were supplied from previous studies of the clear speech intelligibility benefit to normal-hearing listeners in noise. These are reported in Krause & Braida (2002), Bradlow & Bent (2002), and Smiljanic & Bradlow (2005, 2008). We conducted a fifth production study in which 31 native speakers of American English were recorded producing the BKB set of 96 sentences (Bench *et al.*, 1979). We then conducted a study of intelligibility to listeners for 4 talkers selected from the 31 of the production study. Of these 4 talkers, we chose one who we conjectured would produce intelligible Clear Speech, one with likely-unintelligible Clear Speech, and two who we thought were mediocre performers.

TABLE 1. *Characteristics of Datasets Used for Analysis.*

Reference article	# talkers	#listeners per stimulus	Type and # of Stimuli
Bradlow & Bent (2002)	2	10	BKB sentences (64)
Krause & Braida (2002)	5	8	Harvard sentences (105)
Smiljanic & Bradlow (2008)	6	10	Semantically anomalous sentences (20)
Smiljanic & Bradlow (2008)	6	10	Short Paragraphs (2)
Boyce et al. (2011)	4	9	BKB sentences (80)
Boyce et al. (2007, 2011)	31	NONE	BKB sentences (96)

Methods and results

Phase I: The *SpeechMark* measurement system was applied to the five different collections of speech recordings described above (see Table 1). Three measures based on the ***SpeechMark*** system were computed for each data collection : (1) ***Total Number of Landmarks per Sentence***, (2) ***Total Duration of Sentence***, and (3) ***Total Number of Landmark Clusters corresponding to syllabic units in a Sentence*** (i.e. ***Syllabic Clusters***). Measure (3) is a rough measure of syllabic complexity. For this measure, the automatic procedure identifies and tabulates groupings of landmarks that correspond to phonotactically possible syllables of English according to a set of rules that reflect distributional characteristics in the speech signal (Fell et al., 2002).

Table 2 shows the reported mean intelligibility benefit for Clear Speech vs. Conversational Speech in each of the above studies. The intelligibility gain of 13 - 19 percentage points reported by Bradlow & Bent, Krause & Braida, and the two Smiljanic & Bradlow studies is typical of studies on Clear speech for normal-hearing listeners in noise. The lower mean Clear Speech benefit of 6% found in our Boyce *et al.* study reflects the fact that of our four talkers, we deliberately chose three who were likely to produce ineffective or mediocre Clear Speech.

TABLE 2*. Correspondence between Intelligibility and Landmark measures: # of Landmarks, # of Syllabic Clusters, and Total Duration of the Sentence, and intelligibility across all talkers in all databases described in Table 1. All measures show percent change in mean* ***± standard error of the mean (SEM).*** *All measure comparisons for all databases were significant at p < .001 (z test). Data for the four talkers used in the Boyce et al. listener intelligibility experiment are listed separately. *Bradlow & Bent report a mean of 15 RAU, approximately equivalent to a value in the range 13-19%. **Krause & Braida included tests of deliberately slow Clear Speech, included in this mean value.*

Study	Mean gain in Intelligibility (%)	N (pairs)	Gain ± SEM in # Landmarks (%)	Gain ± SEM in # Syllabic Clusters (%)	Gain ± SEM in Total Duration (%)
Bradlow & Bent (2 talkers)	13-19*	128	25±4	31±4	61±10
Krause & Braida (5 talkers)	12-14	50	17±4	24±5	81±14**
Smiljanic & Bradlow (6 talkers)	18	120	21±4	25±6	27±3
Boyce et al. Production Talkers (31 talkers total)	No data	2976	17±1	15±1	17±1
Mean of Above Studies	12.5-18.5		20	24	47
Boyce et al. (4 talkers)	6	80	14±5	14±8	34±14

Phase II: In Phase II of our study, we hypothesized that the talkers who show the greatest intelligibility gains in Clear Speech would produce more Landmarks and more Syllabic Clusters per sentence than any other talkers of the same study. (Note that talkers in the Krause & Braida study were chosen to be particularly good producers of Clear Speech while talkers from the Boyce et al. study were chosen to range from "best" to "worse".) We plotted the two most significant ***SpeechMark*** measures, ***Mean Number of Landmarks*** vs. ***Mean Number of Syllabic Clusters***, against one another for the 17 talkers from all of the data collections for which we have listener intelligibility data. The talkers are divided into "best" vs. "other" groups by their % intelligibility ranking.

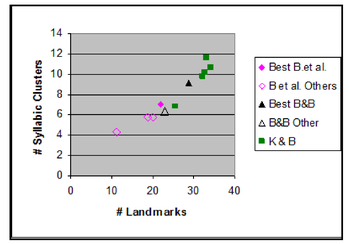

FIGURE 2. *Mean Number of Landmarks vs. Mean Number of Syllabic Clusters for "Best" vs. "Other" talkers across three studies using sentence data (no paragraphs). The B. et al. refers to the 4 talkers of Boyce et al. B & B refers to the 2 talkers of Bradlow & Bent. K & B refers to the 5 talkers of Krause & Braida.*

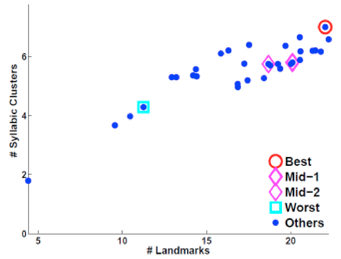

FIGURE 3. *Mean Number of Landmarks vs. Mean Number of Syllabic Clusters for four talkers in Boyce et al, predicted to be best, middling, and worst at using Clear Speech, as compared to other talkers.*

Discussion

Phase I:. Table 2 shows the relationship between the intelligibility data and the Landmark measures as, roughly, the amount of benefit we can assign to each of the Landmark measures ***Mean Number of Landmarks***, ***Mean Number of Syllabic Clusters***, and ***Total Duration***. The fourth row shows the Landmark measures computed from the 27 talkers in the Boyce *et al.* dataset who were not selected for the listener intelligibility experiment. Each Landmark system measure for each study was significantly correlated with listener intelligibility at the $p < .05$ level.

Phase II: Figures 2 and 3 show that ***SpeechMark*** measures parallel this difference in intelligibility. Both figures show separation between "Best" and "Other" talkers; Figure 3 shows that of the 4 talkers in the Boyce et al. study, the predicted "best" and "worst" talkers were most separated, while the predicted "middle" talkers lie in between. The pre-screened Krause & Braida talkers showed the strongest Landmark pattern among the "best" talkers. In contrast, the three talkers from the Boyce et al. study who were selected to produce mediocre or ineffective Clear Speech show the lowest concentration of SpeechMark measures. Thus, this figure is probably representative of the range of variability in Clear Speaking style among people who have not been trained on Clear Speaking style or selected for natural ability.

Conclusion

Our conclusion is that Landmark system measures can reliably detect differences between Clear Speaking style and Conversational Speaking style. Further, these results give us confidence that the Landmark measures provide a reliable and accurate model of effective Clear Speaking style as produced by "best" vs. "worse" talkers.

References

Boyce, S., Fell, H., & J. MacAuslan. (2012). SpeechMark: Landmark Detection Tool for Speech Analysis. In Proceedings of Interspeech, 13th Annual Conference of the International Speech Communication Association 9-13, Portland, OR.
Bradlow, A. & Bent, T. (2002). The clear speech effect for non-native listeners. J. Acoust. Soc. Am. , 112 (1), p. 272-284.
Fell, H. J., MacAuslan, J., Ferrier, L. J.,Worst, S. & Chenausky, K. (2002). Vocalization Age as a Clinical Tool. Proceedings of ICSLP (International Conference on Speech Processing), Denver, USA, September.
Ferguson, S.H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. J. Acoust. Soc. Am. , 116, 2365-2373.
Krause, J.C., & Braida, L.D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. J. Acoust. Soc. Am. ,112, No. 5, pp. 2165-2172.
Krause, J. C, & Braida, L. D. (2004). Acoustic properties of naturally produced Clear speech at normal speaking rates. J. Acoust. Soc. Am. , 115, 362-378.
Liu, S. (1995). Landmark detection in distinctive feature-based speech recognition. M.I.T. Ph.D. dissertation. Cambridge, MA.
Smiljanic, R. & Bradlow, A. (2005). Production and perception of clear speech in Croatian and English. J. Acoust. Soc. Am. , 118, 1677-1688.
Stevens, K.N., Manuel, S.Y., Shattuck-Hufnagel, S., & Liu, S. (1992). Implementation of a model of lexical access based on features. In J.J. Ohala et al. (Eds), *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP), Edmonton: University of Alberta.*
Uchanski, R. M. (2005). Clear speech, In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception*, Blackwell Publishers, Malden, MA, p. 207-235.