# DIFFERENCES AMONG SPEAKERS IN ARTICULATION OF AMERICAN ENGLISH /r/: AN X-RAY MICROBEAM STUDY

*John R. Westbury[1], Michiko Hashi[1], and Mary J. Lindstrom[2]*
*[1]Waisman Center and Department of Communicative Disorders*
*[2]Department of Biostatistics*
*University of Wisconsin-Madison*

## ABSTRACT

Lingual fleshpoint positions and formant frequencies were measured at phonation onset in repetitions of the word *row* from an X-ray microbeam database including 55 normal speakers of American English. These data were used to develop a quantitative description of interspeaker variation in tongue "shape" for /r/, and to determine whether shape variations were acoustically significant, and/or related to gender and variation in selected measures of oral cavity size and shape.

## INTRODUCTION

The x-ray microbeam (XRMB) technique is one of several contemporary methods for studying speech movement. The technique uses a narrow, high-energy x-ray beam (0.4 mm in diameter), controlled by computer, to track the real time motions of small gold pellets (2-3 mm diameter) glued to a speaker's head, lips, tongue, and lower jaw. Thus, the XRMB provides a *point-parameterized* view of speech movement [1], expressed in terms of the time-varying, digitally-sampled positions of discrete articulatory landmarks and fleshpoints.

The XRMB technique is not new. It has been available (albeit on a limited basis) for roughly 20 years, originally at the University of Tokyo where it was first implemented by members of the Research Institute of Logopedics and Phoniatrics [2]; and more recently, at the XRMB facility of the University of Wisconsin (UW) at Madison. The technique was developed as an alternative to high-speed, flood-field cineradiography, and has three significant advantages relative to that method, yielding more accurate data; involving significantly less exposure to ionizing radiation; and, imposing smaller data reduction burdens on the part of those who hope to analyze the information. A natural benefit is that it is now possible to collect and analyze data sets spanning many more speakers and task performances than were feasible using older methods. This "new" development, which we are just beginning to exploit, is important because many physical details of speech production behavior are variable within and across speakers. Generalizations about speech movement, for use in fields such as speech pathology, must be based on samples of speakers and tasks broad enough to reliably reflect the distribution of normal behaviors. Otherwise, there can be no good basis for distinguishing common, ordinary movements made by ordinary speakers, from those that are uncommon and extraordinary.

Over the past five years, a large-scale, freely-available speech production database has been developed at the UW XRMB facility. This database incorporates representations of lingual, labial, and mandibular movements, recorded in association with the sound pressure wave, for more than 50 normal, young adult speakers of American English, for a rich set of utterances and oral motor tasks, and lengthy recording interval (ca. 18 minutes/speaker). The large number of speakers makes this material especially well-suited for analyses of inter-speaker variation in articulatory kinematics.

We have selected a subset of materials from this database to examine production behavior for American English /r/. This sound is interesting and problematic from several points of view. From the acoustic theory of speech production [3], we know that the distinctively low third formant of /r/ can be approximated by vocal tract constrictions in three regions along the vocal tract length. This fact may be related to the kinds and frequency of misarticulations and substitutions that American children make for the sound, and also related to their tendency to master its production late in acquisition [4]. The sound /r/ is also unusual in the kind and degree of variation across major dialects of American English. Moreover, foreign speakers learning the language often find the American /r/ hard to say, and/or hear [5].

Some or all of these facts may be connected -- though precisely how is hard to say -- to an observation made by linguists for many years [6], to the effect that at least two distinct articulatory varieties of /r/ appear to exist, side by side. One broad type is the so-called *retroflex* variety, during which the tongue tip and blade are lifted, and the apex is curled backward in the mouth. The other is the *bunched* variety, where the apex and blade are held low while the front and dorsum of the tongue are elevated. In both varieties, speakers presumably attempt to achieve the same end, forming a primary oral constriction in the mid-palatal region. From a production point of view, /r/ is then interesting because there are different places along the vocal tract where its constriction can be formed, and for one of these places, different ways that the constriction can be formed.

Together, a handful of qualitative, descriptive studies [7-10] have addressed the basic accuracy of the retroflex-bunched distinction; and, only one of these, the remarkable study of Delattre and Freeman [7] of almost 30 years ago, attempted to describe variation in /r/ productions across a large speaker and task sample. We have set out to revisit the topic of /r/ production variability, among a speaker sample somewhat larger than that of Delattre and Freeman, and in so doing, address three specific goals. The first is to develop a quantitative description of variation in tongue posture or "shape" at an acoustically-defined *r-moment* in isolated examples of the word *row* produced by each speaker in the sample. The second goal is to determine whether variation across speakers in tongue shape, at a specific *r-moment,* might be related to variation in formant frequencies at (about) the same moment. And, the third goal is to determine whether the /r/ shapes assumed by speakers' tongues in *row* are related to gender, and/or to selected measures of oral cavity size and shape.

## METHODS

The XRMB speech production database [11] incorporates material from 57 normal, native speakers of American English. Fifty-five (55) of these, 30 females and 25 males, are represented in our analysis of /r/. For this sub-sample, the median age was 21.0 years, with ages ranging between 18.3-37.0 years. For dialect purposes, 29/55 could be considered residents of Wisconsin, while 17 of the remaining 26 were residents of seven neighboring mid-western states of Minnesota, Illinois, Missouri, Iowa, Michigan, Indiana, and Ohio. Dialect homes of the remaining nine speakers were distributed across the US, from Massachusetts (1) to California (2).

Kinematic data recorded from each speaker represent the time-varying, mid-sagittally-projected positions of a set of articulator pellets. For 53/55 speakers, four such pellets were arrayed along the tongue midline. One of these (labelled T1) was always placed in the vicinity of the tongue blade, about 1 cm behind the apex of the extended tongue; a second (labelled T4) always placed in the vicinity of the tongue dorsum, about 6 cm behind the apex; and, two others (labelled T2 and T3) positioned to divide the interval between T1 and T4 into two roughly equal segments. For 2/55 speakers, only three tongue pellets were available. Other articulator pellets were attached to each speaker's mandible, and upper and lower lips.

Pellet-position data were expressed within a rectangular, anatomically-defined coordinate system [12]. The x-axis of the system corresponded to the intersection of each speaker's midsagittal and maxillary occlusal planes. The y-axis was normal to the maxillary occlusal plane (MaxOP), and passed through a local origin at the point where the central maxillary incisors intersected that plane. Thus, *up* in this coordinate system points toward the top of the head along lines perpendicular to the MaxOP; and *forward,* toward the front of the face along lines parallel to the MaxOP, for each speaker.

### Tongue shape measurements for /r/

Pellet positions were tracked at rates ranging between 40-160 times per second, as each speaker read through a list of records containing verbal and oral motor tasks. A subset of five records contained isolated instances of the word *row,* separated in time from different words in the same record, by 0.5-1.0 second. The moment of phonation onset was marked from oscillograms of the acoustic wave recorded during each instance of *row,* articulated by each speaker. Coordinates of all midline tongue pellets were extracted at the time of this event. These coordinates suggest the shape of the tongue at this discrete *r-moment,* and are the focus of our analyses.

In qualitative terms, most speakers prepare to say the /r/ of isolated *row* by drawing some forward part of the tongue up in the mouth, toward the palate, reaching an extreme local configuration some 50-100 ms before phonation onset. Speakers hold this posture for 50 ms or so beyond phonation onset, and then move the tongue rapidly downward, away from the palate, and variably forward or rearward, depending upon speaker and part of the tongue, toward a configuration suitable for the mid-back, diphthongal coda /o$^{u}$/.

Sample data from two speakers are shown in Figure 1. Shapes of the midline tongue contours at phonation onset for /r/ (computed across 4-5 repetitions of the word) are suggested by the average locations of pellets T1-4, connected by solid lines. Ensemble average pellet trajectories are also shown. These indicate paths traced by all four pellets during the interval spanning (-100,+500)ms relative to phonation onset. The pellet locations and trajectories are bounded above by piecewise continuous outlines of each speaker's palatal vault.

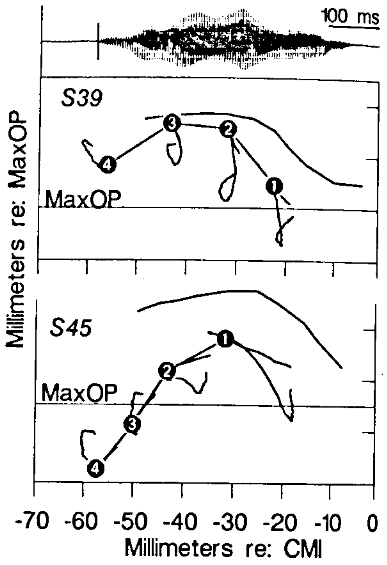

*Figure 1. Above: Oscillogram of one subject's utterance of "row," and tongue contour at /r/ onset (marked by vertical line in the oscillogram). Below: Another subject's contrasting tongue contour.*

Eight coordinates of four pellets provide a sense of tongue shape that is not very tractable. This is so partly because the number of values is high. A simpler expression of these data that reduces their dimensionality, and has the added advantage of emphasizing only tongue shape (and excluding tongue position) for each speaker, has the form of an ordered triple of angles, representing the orientation of straight lines drawn to connect positions of adjacent pairs of pellets. In our data, we designated the orientations of lines connecting pellet pairs {(T1,T2),(T2,T3),(T3,T4)} as angles (1,2,3), respectively. Angle triples (in degrees) for speakers 39 and 45, shown in Figure 1, were (-49,-5,32) and (28,56,50), respectively.

### Acoustic measurements

Formant frequencies were measured from the digitized acoustic waveform, originally sampled at 21.74kHz. LPC and FFT spectra were generated using CSPEECH [13], for an analysis window 20 ms wide, centered at +20 ms with respect to phonation onset, for each token of *row* produced by each speaker. Estimates of formant frequencies from LPC analysis for each token were verified against wide-band spectrograms and corresponding FFT spectra. Bandwidths for spectrograms were 300Hz and 500Hz, for male and female speakers respectively, and the dynamic range was set to 72dB. The number of coefficients for LPC analysis was typically higher than the customary 24, and ranged between 30-40. The higher number of coefficients made certain formant identifications easier, and enhanced our ability to distinguish close second and third formants. Final acoustic measurements for each speaker represented mean formant values calculated across *row* repetitions.

### Oral cavity size and shape

Three indices of oral cavity size and shape were derived from caliper measurements of stone models of each speaker's maxillary dental arch and palatal vault. These indices included mid-sagittal height of the palatal vault (*palht*), above MaxOP, measured 35 mm posterior to the central maxillary incisors; width of the maxillary arch (*m2wid*), measured between distal-buccal cusp tips of the second maxillary molars; and, distance rearward from the central maxillary incisors of the straight line connecting distal-buccal cusp tips of the second maxillary molars (*m2ap*), measured along a line parallel to MaxOP. A fourth index of cavity size -- distance from the central maxillary incisors rearward to the mid-sagittal outline of the posterior pharyngeal wall (*phap*), also measured along MaxOP -- was determined from a calibrated, sagittal-plane x-ray scan of each speaker's oral cavity.

### Statistical methods

Several exploratory statistical techniques were used to gain insight into data collected for this study. All analyses were performed using S-Plus [14] or SAS [15]. The techniques included *hierarchial clustering* [16], to find transformations of the original pellet position data that captured intuitive shape information; *principal component analysis,* to determine the character and explanatory strength of various speaker-by-measure matrices; and *canonical correlation* [17] and *linear regression of ranks,* to search for associations between groups of measurements (e.g., between speaker-by-formant-frequency and speaker-by-tongue-shape matrices). The philosophy underlying data analysis was to capture and describe the extent and nature of variation among measurements of speakers and their articulations, in few terms, without greatly sacrificing interpretability.

Hierarchial clustering, which creates a hierarchy of groups from multi-dimensional data, played a central role in our attempts to describe and understand the pellet-position data. The result of an analysis of this type proceeds from one extreme where every individual is a group of one, to an opposite state in which all individuals form a single group. At each level within trees generated from such analyses, two groups which were closest together in terms of Euclidian distance were combined.

Statistical methods for choosing the "right" number of groups from data arrays do exist, but the methods are not robust. Moreover, they require two assumptions that we were unwilling to make. The first is that some underlying number of groups is already known to exist in the data. The second is that the data in each group follow some pre-specified distribution (e.g., multivariate normal). We chose not to look for some "right" number of clusters in our data, but to use entire hierarchical clustering trees to decide whether multivariate inputs to the procedure (e.g., various transformations of the eight original pellet coordinates) captured shape information that was intuitively salient.

## RESULTS

### Tongue shape

Each speaker made essentially the same lingual gesture during /r/, achieving the same general tongue shape at phonation onset, each time they repeated the word *row*. When we view the entire set of (average) tongue shapes achieved by all speakers, we can readily point out some that look *bunched* (S39 in Figure 1), and others that look *retroflexed* (S45 in Figure 1). But, we also see shapes for some speakers that are not easily matched to these conventional descriptive labels. Two of these that are fairly easy to characterize are the tongue shapes that are relatively flat; and, those that are noticeably "tilted" (retroflexed?) but also somewhat convex (bunched?). Broadly, to the eye, these two shapes seem to be intermediate between those we might categorize as *retroflexed* and *bunched*, and therefore suggest a partition of the data into more than the two simple categories promoted by classical phonetic accounts. However, deciding how many categories there might be, and whether they correspond to the three suggested by Hagiwara [10], or the six suggested by Delattre and Freeman [7], or any other reasonable number (smaller than the speaker sample size!), is difficult. By eye, we can find speaker subsets amongst which roughly the same shape is achieved, though at the same time, it also seems true that these subjectively-identified subgroupings are never exhaustive or mutually exclusive.

More objective partitions, for a number of different expressions of the original pellet-position data, were obtained from hierarchical cluster analysis. However, the outcome of each such analysis was then judged subjectively, so that by eye, groupings found by the analysis had to be confirmed as "reasonable." The success of hierarchical clustering applied to tongue shape data was sensitive to the way the data were expressed. For example, different clusters, different numbers of major shape types, and different principal dimensions in the data were obtained from analyses on (1) the original speaker-by-coordinates array; (2) a similar speaker-mean-centered array (that removed tongue position from the original data); (3) a speaker-by-coefficients array (where the coefficients represented parameters of least-squares quadratic fits to the original coordinates); and, (4) the speaker-by-angles array that we finally found to be most useful. In part, we selected the angle expression of our original data because hierarchical clustering of the data array yielded groupings of speakers by tongue shapes that were intuitively satisfying. In Figure 2, we use scatterplots of tongue shapes to illustrate groups found from "angle" data. We have chosen to plot four groups because these give a visually pleasing result. However, we do not mean to imply that four is the "correct" number of groups.

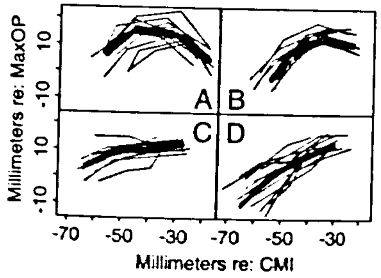

*Figure 2. Scatterplots of tongue shapes showing four groups based on segment angles.*

An angle expression of our data is also useful because it is easy to interpret in a way that directly suggests information about tongue shape. Speakers with negative first angles were those with a blade pellet (T1) that was lower in the mouth than the front-most intermediate pellet T2. Conversely, speakers with a positive first angle were those with the tongue blade (and T1 pellet) higher in the mouth, above MaxOp, than the portion of the tongue represented by the T2 pellet. The second angle was positive for speakers with strongly "tipped" (retroflexed?) tongues, and negative or near zero for those with more convex (bunched?) shapes. The third angle was positive for most speakers, indicating a dorsal pellet (T4) that was lower in the mouth (relative to MaxOP) than all other lingual pellets. The only speakers for whom angle three was not strongly positive were those with relatively flat tongue shapes.

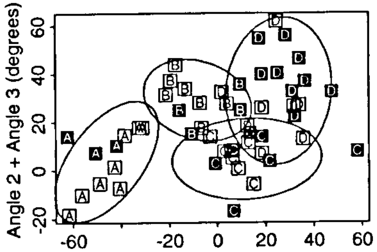

*Figure 3. Tongue shapes plotted on an approximate principal component plane.*

The first two principal components from an analysis of the speakers-by-angles data array explained 95% of the variability. This fact allows us to display tongue shape information, expressed in terms of angles, in two dimensions without a major loss of information. The first principal component was dominated by the first angle (defined by T1 and T2 pellet positions), while the second principal component was dominated by the sum of the second and third angles. The distribution of speakers' tongue shapes, expressed in terms of angles and plotted in the plane defined by these two approximate principal components, is illustrated in Figure 3. The approximate principal components are more easily interpreted than the actual components defined in analysis, and still explain 89% of the variability across speakers. Male and female speakers are distinguished by filled and open squares. Categories of four high-level shapes uncovered by hierarchical clustering are coded by letter, and two-sigma ellipses are drawn about coordinates of each category's centroid. Perhaps the simplest lesson that Figure 3 teaches is that the range of tongue postures for /r/, viewed in this way, is more nearly continuous than categorical across speakers, along either approximate principal component dimension. Simply put, this expression of our data seems to argue against discrete types of tongue shapes for /r/.

### Formant frequencies

*Table 1. Mean formant frequencies (in Hz) for male and female speakers. Standard deviations are in parentheses.*

| | Male | Female |
|---|---|---|
| F1 | 326 (39) | 358 (45) |
| F2 | 882 (90) | 1092 (122) |
| F3 | 1378 (121) | 1792 (201) |

Expected differences were found between formant frequencies for male and female talkers. Average frequencies for F1 and F2 agreed well with comparable data from normal geriatric speakers[18].

Interestingly, no significant association was found between speaker-by-tongue-shape and speaker-by-(log-transformed)-formant-frequency arrays. Thus, it appears that the very large differences across speakers in tongue shape at phonation onset in *row* do not seem to be accompanied by statistically reliable differences formant frequencies.

### Oral cavity size

*Table 2. Mean measures of oral cavity size (in mm), for males and females. Standard deviations are in parentheses.*

| | Male | Female |
|---|---|---|
| *palht* | 22.4 (2.2) | 18.7 (2.2) |
| *m2wid* | 60.3 (4.7) | 57.0 (3.1) |
| *m2ap* | 43.7 (3.1) | 41.5 (3.8) |
| *phap* | 80.2 (5.0) | 77.7 (4.1) |

On average, males were slightly larger than females for all measures of oral cavity size. Male palates (*palht*) were about 3mm taller, and male maxillary arches at the second molar tooth (*m2wid*) were about 3 mm wider. Gender and oral cavity size are therefore confounded variables across our speaker sample. Across all speakers, the highest palatal vault was about 26 mm, while the shallowest was only 14 mm. The widest arch was 68 mm, while the narrowest was only 48 mm.

The only statistically significant relationship between our speaker-by-angle characterization of tongue shape for /r/, and measures of oral cavity size and shape, and gender, was between the first angle and gender. This effect is suggested in Figure 3, in which females, as a class, seem to have more negative first angles than do males. However, this effect was not strong. Gender explained only a small proportion of the variability in the first angle across speakers ($r^2$ = 0.13). Moreover, the association between the first angle (defined primarily by relative height of the tongue blade), and gender, has not been found in preliminary analyses of tongue shapes for /r/ in the words *street*, *problem*, *right*, and *across*.

## DISCUSSION

How many kinds of tongue shapes exist for /r/ in American English? This simple question, asked by others before us, presumes that data should segregate into a number of discrete articulatory categories. However, our data seem to argue against such an assumption. No matter what visual and numeric tricks we have tried, the data summarized in this report do not seem to distribute well into discrete categories. It is probably closer to the truth that there is a continuous range of acceptable/possible tongue shapes for /r/. Speakers must achieve an acoustic result their listeners will accept as /r/. Precisely how that result is obtained, using the tongue and lips to constrict the vocal tract tube near either end, and/or near its middle, may be physiologically important to individuals, but not in a way that forces different speakers to achieve an invariant articulatory result. This is not a new idea, though tongue shape data for /r/, collected across many speakers, illustrate the idea perhaps more vividly than other data types.

The opportunity to examine data from many speakers is a main benefit of the XRMB method and database. Such an opportunity is necessary if we hope to know the distribution of tongue shapes for /r/ that exist in American English. This information is theoretically interesting, and may also have some practical benefit for speech therapy. Speakers who do not produce acceptable variants of /r/ are coached by therapists to achieve better results through instructions expressed in articulatory terms: instructions to shape the tongue in some particular way, and/or to place the tongue in some specific location. Speakers who are informed of the range of known and possible articulatory options may then choose some optimal variant.

The fact that variation across speakers in tongue shapes and formant frequencies at the same *r-moment* in *row* are not well related is superficially surprising. Differences between some speakers' tongue shapes, for the moment we have examined, are extreme. However, we can excuse the lack of relationship in view of standard acoustic theory [3]. The shape of the radiated spectrum depends heavily upon the vocal tract area function, and the area function itself is only partly determined by tongue shape within the oral cavity. The degree and locations of all constrictions along the vocal tract length define the area function. For /r/, constrictions in the pharynx, and at the lips, may be especially important. In our data, the former is inaccessible. The latter is somewhat less so, though the information available, given by positions of pellets on the lips, is difficult to interpret. Even the information we have for the oral portion of the tongue is less complete than we might like. Of course, the position of the apex is lost from our data, and this loss may be a special problem for any attempt to understand the

articulation of /r/. We also have only a coarse outline of the tongue, defined by fleshpoints locations in the vicinity of the blade and dorsum, and two points in between. In principle, we might still expect some relationship between articulation and the acoustics of /r/, though for our data, we also have many reasons to reject that expectation.

The fact that variation across speakers in tongue shapes for /r/, and size and shape of the oral cavity, were not strongly related is also something of a surprise. We often assume that how speakers move when they speak depends partly upon how they are built. Our data for /r/ productions seem to show that this is not true, though it important again to emphasize the coarse nature of postural and size data. Certainly from our data, we cannot yet suggest why speakers choose the shapes they do for /r/.

## ACKNOWLEDGEMENT

Research supported by USPHS Grant DC00820, and a collaborative research agreement between the University of Wisconsin-Madison, and ATR Human Information Processing Research Laboratories.

## REFERENCES

[1] Houde, R. A. (1967), "A study of tongue body motion during selected speech sounds", Ph.D. dissertation, University of Michigan.

[2] Fujimura, O., Kiritani, S. & Ishida, H. (1973), "Computer-controlled radiography for observation of movements of articulatory and other human organs", Comput. Biol. Med., vol. 3, pp. 371-384.

[3] Fant, G. (1980), "The relations between area functions and the acoustic signal", Phonetica, vol. 37, pp.55-86.

[4] Shriberg, L. D. (1993), "Four new speech and prosody-voice measures for genetic research and other studies in developmental phonological disorders", Journal of Speech and Hearing Research, vol. 36, pp.105-140.

[5] Yamada, R. A. & Tohkura, Y. (1992), "Perception of American English /r/ and /l/ by native speakers of Japanese", in Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka eds, *Speech perception, production & linguistic structure*, Burke VA: IOS Press, Inc.

[6] Hockett, C. F. (1958), *A course in modern linguistics*, NY: Macmillan.

[7] Delattre, P. & Freeman, D. C. (1968), "A dialect study of American R's by X-ray motion picture", Linguistics, vol. 44, pp. 29-68.

[8] Zawadzki, P. A. & Kuehn, D. P. (1980), "A cineradiographic study of static and dynamic aspects of American English /r/", Phonetica, vol. 37, pp. 253-266.

[9] Lindau, M. (1985), "The story of /r/*", in V. A. Fromkin ed, *Phonetic linguistics, essays in honor of Peter Ladefoged*, NY: Academic Press.

[10] Hagiwara, R. (1994), "Three types of American /r/", UCLA working papers in phonetics, vol. 88, pp. 51-61.

[11] Westbury, J. R. (1994), *X-ray microbeam speech production database user's handbook, version 1.0*, Madison WI.

[12] Westbury, J. R. (1994), "On coordinate systems and the representation of articulatory movements", Journal of Acoustical Society of America, vol. 95, pp. 2271-2273.

[13] Milenkovic, P. & Read, C. (1992), *CSpeech Version 4 User's Manual*, Madison WI

[14] Becker, R. A., Chambers, J. M., Wilks, A. R. (1988), *The new S language: a programming environment for data analysis and graphics*, Belmont, CA: Wadworth.

[15] SAS Institute Inc. (1989), *SAS/STAT user's guide, version 6*, Carey, NC: SAS Institute Inc.

[16] Everitt, B. (1980), *Cluster analysis*, NY: Halsted.

[17] Dillon, W. R. & Goldstein, M. (1984), *Multivariate analysis, methods and applications*, NY: Wiley.

[18] Weismer, G., Kent, R. D., Hodge, M. & Martin, R. (1988), "The acoustic signature for intelligibility test words", Journal of Acoustical Society of America, vol. 84, pp. 1281-1291.