Phonetica 1988; 45: 30–42

0031–8388/88/0452–0030 $ 2.75/0

A Multidimensional Scaling Study of Esophageal Vowels

Robert Allen Fox, Michael D. Trudeau

Speech and Hearing Science, Ohio State University, Columbus, Ohio, USA

Abstract. A multidimensional scaling experiment was conducted to determine the perceptual structure of 11 American English vowels produced by a competent esophageal speaker. Estimates of perceptual distance among these vowels were obtained using a 9-point similarity/dissimilarity scale and were analyzed using an individual differences scaling algorithm (INDSCAL). A three-dimensional perceptual space was produced. The three perceptual dimensions corresponded to tongue advancement, vowel height, and rhotacization. These three dimensions were then correlated with selected bark scale transformed acoustic vowel measurements. The perceptual dimensions 1–3 corresponded most closely to F_3–F_2, F_1–F_0, and F_3, respectively. Little difference was found between the perceptual structure of esophageal vowels and laryngeal vowels, although it is suggested that the correlation between some of the acoustic measures (such as F_0) and the perceptual dimensions may change as a function of individual speaker differences due to postsurgical capabilities (anatomical and physiological) and/or method of voice restoration.

Multidimensional scaling (MDS) has been used in a number of different studies to determine the salient features or dimensions used in the perception of normally articulated vowels [Fox, 1982, 1983, 1985a; Goldstein, 1971; Pols et al., 1969; Rakerd and Verbrugge, 1985; Shepard, 1972; Singh and Woods, 1971; Terbeek, 1977]. These perceptual dimensions can frequently be related to various phonological features of the vowel stimuli (e.g., distinctive features) and support the importance of such features to the vowel identification process. The perceptual dimensions obtained in various MDS studies have been remarkably stable in their interpretations and have, in general, corresponded to traditional articulatory or acoustic distinctions. Most studies have found two dimensions corresponding to tongue advancement and vowel height distinctions [Pols et al., 1969; Shepard, 1972; Fox, 1982, 1983; Rakerd and Verbrugge, 1985]. Additional dimensions have differed somewhat, for example, Fox [1982] obtained a rounding dimension, Singh and Woods [1971] found a retroflex dimension,

while Rakerd and Verbrugge's [1985] third dimension corresponded to tenseness. The fact that these studies utilized slightly different stimulus sets almost certainly contributed to the obtained dimensional differences.

However, this technique, to our knowledge, has never been used to determine the perceptual features underlying the perception of any type of alaryngeal vowels. Certainly the perceptual features involved in the perception of consonants produced by alaryngeal speakers may differ from those produced by normal, laryngeal speakers. Compare, for example the consonantal perceptual features obtained for esophageal and tracheoesophageal speech by Doyle and Danhauer [1986] with those obtained by Singh et al. [1972] for normal speech.

Although vowels are normally described in terms of articulatory tract configurations (i.e., filter characteristics), it is clear that changes in source characteristics can affect the perception of vowel quality. For example, several studies [Miller, 1953; Slawson, 1968; Fujisaki and Kawashima, 1968; Fox, 1985b] have shown that a given formant pattern may be perceived as representing different vowel qualities depending on the vowel's fundamental frequency (F_0). While no study has shown that source characteristic differences can affect the basic nature of the perceptual dimensions, esophageal speech differs from normal laryngeal speech in more than just voicing source characteristics. Total laryngectomy alters not only the phonatory source for speech, but the configuration of the vocal tract as well [Diedrich and Youngstrom, 1966]. The observation that the midpoint frequencies of the first three formants (F_1, F_2, F_3) in esophageal speech are elevated in comparison to the formant structure of normal laryngeal speech has been attributed to the altered physical configuration of the vocal tract [Sisty and Weinberg, 1972]. In that report, Sisty and Weinberg [1972, p. 445] also noted that the F_1 by F_2 vowel plots for esophageal speech were dissimilar to the analogous plots for laryngeal speech and that among the vowels sampled some demonstrated 'an unexpectedly marked reduction in the intensity of the third formant'. It is unknown whether or how these acoustical differences affect the perceptual dimensions listeners utilized in discriminating among vowels in esophageal speech.

The present study uses multidimensional scaling (utilizing the program INDSCAL) to examine the perceptual dimensions underlying the perception of a representative set of esophageal vowels produced by a single competent speaker and compares these dimensions with those previously found in MDS studies utilizing normally voiced vowels.

Method

Speaker

The esophageal speaker in this investigation was a 73-year-old, college-educated female who had had a total laryngectomy 9 years and 8 months prior to the investigation. Esophageal speech was her sole means of vocal communication. Although her intelligibility was not formally assessed, the subject was a speaker for the American Cancer Society and frequently addressed local high schools, professional groups and nursing training programs. The speaker's level of acceptability in esophageal speech had been determined in an earlier investigation [Trudeau, 1987] in which she demonstrated the median in speech acceptability in a group of 5 good to excellent female esophageal speakers. For inclusion in that study, all speakers were prescreened by the second author as at least average in esophageal

speech proficiency. Following this, a panel of 3 speech-language pathologists, experienced in alaryngeal speech rehabilitation but unaware of the nature of the study or of the subjects' identities, assessed each subject's proficiency in esophageal speech. This was assessed on a 0–3 ordinal scale with 0 as below average and 3 as superior. The subject in the current study was rated at 2 (above average) by all 3 clinicians. Next, 24 naive listeners were trained in the use of a 5-point equal-appearing interval scale [Thurstone and Chave, 1929] for measuring acceptability of alaryngeal speech. For this task and for the experienced listeners' judgments, the second and third sentences from the 'Rainbow Passage' [Fairbanks, 1960] served as the stimulus. The 24 listeners' ratings for esophageal speech produced by 5 females ranged from 1.64 to 2.88. The present subject's rating of 1.92 was the median. Based on this evaluation process we felt that the present speaker exhibited above-average, but not excellent, proficiency in esophageal speech and, therefore, represented speech typical of the successfully rehabilitated esophageal speaker.

Recording Procedures

This speaker recorded the reading passage and the vowels [i ɪ ɛ æ ɑ ʌ oʊ ʊ u ɔ ɚ] while seated in an audiometric booth wearing a headset-mounted microphone to insure a constant mouth-to-microphone distance of 10 cm. This vowel set was representative of the relatively monophthongal vowels found in Midwestern American English, which was the speaker's dialect. The vowel set included 4 front vowels [i ɪ ɛ æ], 3 central vowels [ɑ ʌ ɚ], and 4 back vowels [u ʊ oʊ ɔ]. The somewhat diphthongal vowel [oʊ] was used to ensure this more balanced pattern. In addition to the microphone, equipment included a high-quality stereo cassette deck and high bias audio cassettes.

The speaker was provided with the 98-word version of the 'Rainbow Passage' and the 11 vowels in the context of /h____d/with the carrier phrase of 'I will say ____' [Peterson and Barney, 1952]. This format was followed, despite the inability of the speaker to produce the glottal fricative in esophageal speech, to provide a constant articulatory context. The speaker was allowed to familiarize herself with these materials and to practice reading them aloud prior to recording. She was also encouraged to rerecord the passage or any sentence which she or one of the investigators felt she did not produce accurately.

Stimulus Preparation

The sentences produced by the speaker were low-pass-filtered at 4.5 kHz, digitally sampled at 10 kHz and stored on a computer disk. The test vowels were then edited from these sentences using the ILS waveform editor (Signal Technology, Inc.). In all cases the onset of the target vowel was easily identifiable in both the oscillographic trace (used in waveform editing) and in corresponding spectrographs. Care was taken to eliminate all final consonant transitions. The mean amplitudes of these vowel tokens were then equalized within a 3-dB range. The generation of the stimulus tape, using these sampled tokens, was done under computer control. The vowel tokens appeared in pairs, 300 ms apart on the stimulus tape. Only one example of each vowel token was used in the creation of the stimulus tape. The interval between separate stimulus pairs was 4 s.

With 11 different tokens, there were 110 different vowel pairs possible, not allowing stimulus tokens to be paired with themselves. Two different blocks of 110 vowel pairs (separately randomized) were produced. Each listener thus made a similarity judgment on 220 vowel pairs.

To determine the extent to which these excised vowels, in fact, had the appropriate (intended) vowel quality, a listening test was conducted using 10 listeners, each of whom was a graduate student at Ohio State (none of these listeners participated in the scaling task itself). Each of the 11 vowels was presented 15 times (producing 150 vowel identifications) in random order to the listeners. There was a 5-second pause between vowels. Listeners were required to indicate which vowel they heard by circling the word on the response sheet which contained that vowel (each line of the response sheet contained the words *heed, hid, head, had, hod, hud, hawed, hoed, hood, who'd, heard).* The vowels [i ɛ æ oʊ ʊ u and ɚ] were identified as their intended vowel qualities over 90% of the time. The vowels [ɪ and ʌ] were identified as [ɪ] and [ʌ] 86 and 84% of the time, respectively. The vowel [ɑ] was identified as [ɑ] 55% of the time, but was often confused with [ɔ] (43% of the time). The vowel intended as [ɔ] was only identified as [ɔ] 20% of the time and was most often identified as [oʊ]. These confusions were not

Table 1. Measured acoustic parameters for stimulus vowels (in hertz)

	F_1			F_2			F_3			F_0	Duration
	onset	mid	offset	onset	mid	offset	onset	mid	offset		
i	374	406	343	2,917	2,688	2,724	4,095	4,078	3,937	74	459
ɪ	422	422	457	2,706	2,601	2,442	4,060	3,954	3,936	74	272
ɛ	566	546	508	2,284	2,285	2,246	2,929	2,910	2,831	59	272
æ	721	738	756	2,214	2,214	2,003	2,900	2,847	2,653	55	469
ɑ	791	738	756	1,212	1,142	1,282	2,654	2,636	2,618	60	433
ʌ	773	738	773	1,423	1,300	1,458	2,689	2,654	2,706	70	259
oʊ	703	526	474	1,001	843	896	2,724	2,496	2,584	80	351
ʊ	562	562	615	1,019	1,248	1,353	2,284	2,530	2,724	84	261
u	369	369	334	984	878	949	2,494	2,144	2,091	100	310
ɔ	598	615	615	1,072	1,002	914	2,478	2,478	2,513	77	310
ɚ	633	580	615	1,230	1,160	1,195	2,109	1,898	1,916	60	335

unexpected since the Midwestern American dialect of English does not clearly distinguish the vowel [ɔ] from the other back vowels [similar patterns of confusion can be seen in Peterson and Barney, 1952]. These data demonstrate that the stimulus vowels used in this study do, in general, represent the indicated vowel qualities.

Acoustical Analysis of Vowels

So that the acoustic correlates of the obtained perceptual dimensions could be determined, a number of acoustic parameters were measured for each of the 11 vowels. The parameters included the frequencies of the first three formants (at three different points: onset, center, and offset), F_0, and duration. These measurements were made using a digital spectrograph (Voice Identification, model RT-1000). These values are shown in table 1. Formant values were estimated by identifying the top and bottom of the frequency band and taking the mean. The data from Sisty and Weinberg [1972] were used as reference source to aid in the interpretation of the spectrographic displays. F_0 was determined by counting voicing striations evident on the spectrogram. The spectrograph was set in 'zoom mode' allowing isolation and display of each individual vowel. One of the investigators computed these data. A doctoral research assistant, unfamiliar with the data or this study, applied the same procedure in a separate analysis of the same vowels. The Pearson product-moment correlation from the two sets of F_0 measures was 0.96 ($p < 0.001$).

Listeners

There were 20 American English listeners who were relatively unfamiliar with esophageal speech. All listeners were undergraduates at the Ohio State University and were paid $2.50 for their participation. None had any known hearing impairment.

Procedure

This study required listeners to hear two sequentially presented stimulus tokens (a stimulus pair) and to judge their similarity/dissimilarity on a 9-point scale [Fox, 1983]. Listeners indicated their judgments on the prepared response sheets by circling the appropriate point on the rating scale. They were instructed (orally and in writing) to ignore all but vowel quality differences between the two vowels in each trial. Listeners were also cautioned to use the entire range of the 9-point scale. Following these instructions, listeners heard a recorded list of all 11 stimulus vowels twice. Subjects then practiced on an introductory block of 10 stimulus trials. After determining that the listeners were indicating their responses in the correct fashion, the experimenters answered any questions that listeners might have had.

Each listener's similarity judgments were checked for consistency before his data was in-

cluded in the multidimensional scaling analysis. This was done to ensure that excess 'noise' was not included in the perceptual distance data submitted to INDSCAL analysis. As Terbeek [1977] cautioned, excessive noise in the data may cause a solution of too many dimensions to appear more formally acceptable (using various criteria) than a (more correct) solution of fewer dimensions. This consistency check involved comparing a listener's responses on one block of the stimulus trials with his or her responses on the other block using simple correlational statistics. A listener's data were used only if Pearsons's r was 0.50 or greater, ensuring a significance level of at least 0.01. Three listeners failed to demonstrate the necessary consistency and their data were eliminated from all the following analyses. Thus the data from 17 listeners were submitted to INDSCAL analysis.

MDS Analysis

The perceptual data were analyzed using the metric MDS procedure INDSCAL [Carroll and Chang, 1970]. Like most MDS procedures, this technique involves the assumption that the scaled similarity judgments reflect the 'perceptual distance' between the two objects (vowels in this case). MDS algorithms attempt to account for these interobject distance estimates by modeling a perceptual space, complete with coordinate values for the objects within this space. In this study, the vowels are located in the n-dimensional space such that the distance between the objects corresponds optimally to the experimentally obtained perceptual distance estimates. No assumption as to the number or nature of these dimensions is made prior to MDS analysis.

INDSCAL is a powerful MDS analysis program which includes individual differences between listeners in developing its n-dimensional solution and assigns the relative salience (or weight) of each separate dimension for each subject. One of the advantages that accrues as a function of this weighted analysis is that the orientation of the solution is fixed and cannot be rotated without worsening the overall fit of the solution to the data. Thus, problems of solution rotation commonly encountered in factor analysis and older MDS methods are avoided [see discussions in Carroll and Chang, 1970; Kruskal and Wish, 1978; Terbeek and Harshman, 1971].

Solutions were obtained in 1–5 dimensions. Four different solutions were obtained at each dimensionality and each solution had a different random starting configuration to avoid the problem of 'local minima'. In particular, INDSCAL may sometimes find a solution which accounts for a large amount of the variance of the data, but which may not represent the 'best' solution at that dimensionality (the optimal solution at a particular dimensionality, with the least amount of variance unaccounted for, is considered the 'global minimum'). Obtaining several solutions at each dimensionality reduces the likelihood of missing the 'best' solution.

Results and Discussion

INDSCAL Analysis

The three-dimensional solution was chosen as the best solution. The perceptual dimensions for this solution are shown in figures 1 and 2. This decision was based upon three criteria: variance accounted for, uniqueness of solution, and interpretability. In terms of the first criterion, the one-, two-, three-, four-, and five-dimensional solutions accounted for 47, 61, 68, 72 and 77% of the variance, respectively. Basing the dimensionality decision strictly upon variance accounted for, one might have selected dimensionality 2 – however, we feel that the other two criteria dictate that the three-dimensional solution be selected.

The second criterion used was what has been called the uniqueness-of-solution property [Terbeek, 1977]. INDSCAL theoretically provides a unique spatial solution up to the correct number of dimensions and multiple solutions at dimensionalities above the correct dimensionality (each separate analysis produced with a different random starting configuration). For these data, multiple solutions (i.e., solutions whose perceptual dimensions were not directly comparable) were obtained at the four-dimen-

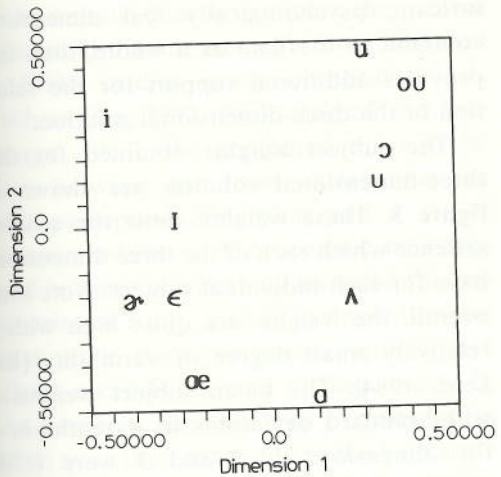

Fig. 1. Dimension 1 by dimension 2 plot of the three-dimensional perceptual space.

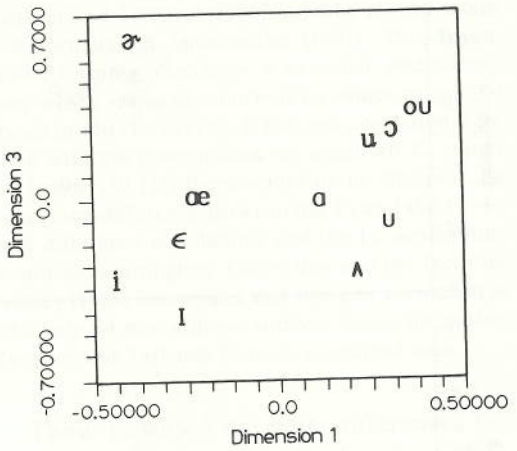

Fig. 2. Dimension 1 by dimension 3 plot of the three-dimensional perceptual space.

sional level. These multiple solutions were not simply the case of finding different local minima because each different solution accounted for the same proportion of the variance (within 0.005%).

The third criterion involved interpretability, which is perhaps the most commonly utilized criterion in MDS studies of vowels. The first three dimensions obtained, in the order of their appearance, could easily be labeled tongue advancement, vowel height, and rhotacization. In addition, the first two dimensions underwent little change in the three-dimensional solution. On the other hand, the four-dimensional solutions significantly disrupted both the tongue advancement and vowel height dimensions and produced dimensions which did not reflect any easily discernible linguistic feature.

As a final check on the correctness of selecting the three-dimensional solution, a version of Gandour and Harshman's [1978] 'split-half' procedure was done. The perceptual data were divided into two separate halves. Each half contained the perceptual distance judgments based on the responses to only one of the blocks of stimuli from each of the 17 listeners. One 'split-half' data set contained the distances obtained from 9 listeners on block 1 of the stimuli and from 8 listeners on block 2 of the stimuli. The second 'split-half' data set contained the distances obtained from those same 9 listeners on block 2 of the stimuli and the 8 remaining listeners on block 1 of the stimuli. Each 'split-half' data set thus represented a different set of perceptual responses from each of the 17 listeners. Each of these two data sets were then analyzed using INDSCAL analysis at three dimensions. Both resulting solutions contained dimen-

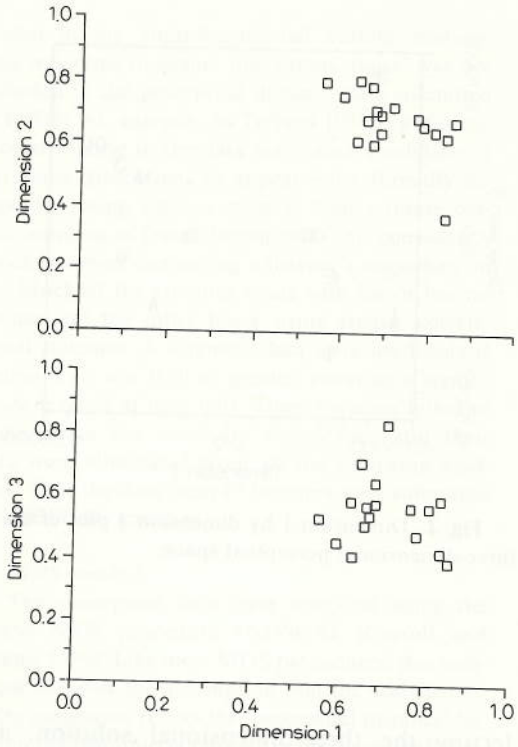

Fig. 3. Subject weight space for the three-dimensional solution.

sions which were similar to the ones obtained using the entire data set and which could easily be labeled tongue advancement, vowel height, and rhotacization. Pearson's correlations were then obtained between the matching dimensions in each solution. The obtained correlation coefficients were very high: dimension 1, tongue advancement [$r(11) = 0.994$, $p < 0.001$]; dimension 2, vowel height [$r(11) = 0.980$, $p < 0.001$], and dimension 3, rhotacization [$r(11) = 0.948$, $p < 0.001$]. If we assume that those perceptual dimensions common to both halves of the data will include the significant, psychologically real dimensions common to the data as a whole, then this provides additional support for the selection of the three-dimensional solution.

The subject weights obtained for this three-dimensional solution are shown in figure 3. These weights show the relative salience which each of the three dimensions have for each individual subject. Note that, overall, the weights are quite high with a relatively small degree of variability [Rakerd, 1984]. The mean subject weights – with standard deviations in parentheses – for dimensions 1, 2 and 3 were 0.709 (0.090), 0.671 (0.024), and 0.536 (0.085), respectively.

Interpretation of the Dimensions

As is evident from an inspection of figures 1 and 2, the three perceptual dimensions can easily be given the traditional linguistic labels of tongue advancement, vowel height, and rhotacization (this third dimension could be termed retroflexion, but we prefer the acoustic term because a rhotic vowel can be produced by articulatory configurations other than retroflexion) [Ladefoged, 1982]. It is important to note that these perceptual dimensions (1) reflect traditional linguistic distinctions associated with vowels in American English and (2) are completely compatible with dimensions obtained in studies using laryngeal speakers and similar sets of vowel stimuli [Singh et al., 1972; Shepard, 1972; Fox, 1982].

Labeling the dimensions, however, gives only a limited, qualitative view of their nature. A better approach is to compare the dimensions to the acoustic characteristics of the vowels themselves. This will allow us to evaluate the relationship between the perceptual and acoustic structures of the stim-

uli and more directly compare the present perceptual dimensions with those obtained in MDS studies using laryngeal vowels [such as Fox, 1982, 1983; Rakerd and Verbrugge, 1985]. We computed the correlations between the vowel coordinates on each perceptual dimension and a set of acoustic measures. The acoustic measures used in these correlations included the frequencies of F_1, F_2, and F_3 at three different locations in the vowel (onset, middle, and offset), mean F_1, F_2, and F_3 frequencies, F_0 and duration. The measured frequencies (in hertz) were then converted to the bark scale [Zwicker and Terhardt, 1980] before correlations were computed.

The physical frequency measures (in hertz) were converted to the bark scale because several researchers [including Bladon et al., 1984; Syrdal, 1985; Syrdal and Gopal, 1986] have argued that formant frequencies in hertz do not accurately reflect the auditory representation of the vowel in the perceptual system. In addition, conversion of the data to the bark scale seems to reduce the vowel normalization problem. See Pols et al. [1969] for a different approach to such comparisons using 1/3-octave frequency band measurements. It should be noted that some researchers [including Nearey, 1988] suggest that there is little evidence favoring auditory scales (mels, barks) over more traditional scales in normalization schemes.

This transformation converted a frequency value in kilohertz to a critical band value in bark using the following formula:

$$B = 13 \arctan(0.76 f) + 3.5 \arctan(f/7.5)^2$$

where B is the critical band in bark and f is the frequency in kilohertz. In addition, several other acoustic measures were computed which may also play a role in a listener's perceptual decisions [Syrdal and Gopal, 1986].

Syrdal and Gopal [1986], rather than using the Zwicker and Terhardt [1980] formula, use the modified formula of Traunmüller [1981]. The Traunmüller formula introduces a so-called 'end correction' which serves to adjust all F_0 values below 150 Hz up to 150 Hz. Clearly, if this end correction were done with the present data (in which all F_0 values are below 150 Hz) there would be no theoretically interesting difference between the F_1 and the F_1-F_0 bark difference correlations and the F_0 correlations would be meaningless. Given this and the fact that Nearey [1988] has argued that this end correction is relatively ad hoc with no auditory basis, the earlier Zwicker and Terhardt formula is utilized here.

These included the bark differences between F_1 and mean F_0 (F_1-F_0), F_1 and F_2 (F_2-F_1), and F_2 and F_3 (F_3-F_2), calculated at the three different locations and their combined mean. Selected correlations are shown in table 2.

The first perceptual dimension (tongue advancement) correlates well with the F_2 and F_2-F_1 bark difference measures but the highest correlations are obtained with the F_3-F_2 bark difference measures. These results are entirely consistent with MDS studies using laryngeal vowels (Fox, 1982, 1983, 1985a; Rakerd and Verbrugge, 1985]. The high correlation between dimension 1 and the F_3-F_2 bark difference measure supports the claim by Syrdal and Gopal [1986] that this measure more easily distinguishes the front/back differences in American English vowels than does F_2-F_1 [suggested by Ladefoged, 1982].

The second perceptual dimension (vowel height) is very highly correlated with the obtained F_1 and F_1-F_0 measures. This is, again, consistent with the other MDS studies cited above. It is important to note that the F_1-F_0 bark difference measure is strongly correlated with dimension 2 [consistent with Syrdal and Gopal, 1986] despite

Table 2. Pearson's correlations between acoustic parameters of the vowel stimuli (in bark) and the obtained perceptual dimensions

Acoustic parameter	Dimension 1	Dimension 2	Dimension 3
Mean F_1	0.23	−0.78**	0.13
F_1 onset	0.30	−0.65*	0.25
F_1 mid	0.19	−0.80**	0.08
F_1 offset	0.17	−0.82***	0.05
Mean F_2	−0.69**	−0.19	−0.79**
F_2 onset	−0.81**	−0.27	−0.67*
F_2 mid	−0.79**	−0.31	−0.71**
F_2 offset	−0.78**	−0.34	−0.76**
Mean F_3	−0.42	0.07	−0.84***
F_3 onset	−0.48	0.11	−0.75**
F_3 mid	−0.42	0.02	−0.85***
F_3 offset	−0.34	0.06	−0.86***
Mean F_1-F_0	0.16	−0.81***	0.11
F_1-F_0 onset	0.23	−0.70**	0.23
F_1-F_0 mid	0.12	−0.83***	0.07
F_1-F_0 offset	0.12	−0.85***	0.04
Mean F_2-F_1	−0.63*	0.24	−0.66*
F_2-F_1 onset	−0.78**	0.03	−0.61*
F_2-F_1 mid	−0.75**	0.05	−0.64*
F_2-F_1 offset	−0.74**	0.08	−0.67*
Mean F_3-F_2	0.86***	0.48	0.51
F_3-F_2 onset	0.89***	0.47	0.49
F_3-F_2 mid	0.88***	0.47	0.45
F_3-F_2 offset	0.87***	0.52	0.48
F_0	0.57*	0.88***	0.06
Duration	−0.34	−0.22	0.09

*$p<0.05$, **$p<0.01$, ***$p<0.001$.

the presence of an alaryngeal voicing source. The F_1 by F_2 plot (shown in fig. 1) is very similar to those obtained in the relevant literature [Singh and Woods, 1971; Fox, 1983; Rakerd and Verbrugge, 1985] with some perturbation evident in the expected positions of a few vowels. For example, one might have expected the vowels [i] and [ɪ] to be somewhat higher compared to the relative position of [u]. One could speculate that had the diphthong [eɪ] (which can be considered the front, unrounded version of [oʊ]) been included in the stimulus set the relative positions of [i] and [ɪ] may have been higher. Rakerd and Verbrugge [1985] also used a stimulus set which included [oʊ] but which did not include [eɪ]. Their F_1 by F_2 plot did not show [i] and [ɪ] to be lower than expected in the perceptual space, although [ɪ] was somewhat higher than [i] on the vowel height dimension. However, some degree of perturbation in the expected position of individual vowels (fine-structure variation) is commonly encountered in perceptual scaling studies as any quick examination of the literature will show. It is the overall pattern and the nature of the perceptual dimensions which are of greatest importance in the present study and from this point of view the perceptual dimensions obtained using alaryngeal vowels are entirely compatible with those obtained using laryngeal vowels.

One result which is not consistent with previous MDS studies is the discovery of a very significant correlation (in fact, the highest correlation) between dimension 2 and F_0. In one sense, a high correlation is not completely unexpected. House and Fairbanks [1953] noted that F_0 was correlated with vowel quality (at least with laryngeal speech), with high vowels tending to have a higher F_0 than low vowels. Gandour and Weinberg [1980] found the same type of relationship between vowel height and F_0 in a group of 4 esophageal speakers.

One common explanation for this relationship in laryngeal speech is the tongue-pull hypothesis [Ladefoged, 1968; Lehiste, 1970]. This hypothesis suggests that the in-

creased tension of the vocal folds is due to the extralaryngeal muscle activity in moving the tongue to a higher, more constricted articulatory position. When the tongue is elevated, a pull is exerted on the larynx, which increases the tension of the vocal folds. It is unclear whether a similar explanation could account for this relationship in esophageal speech. Gandour and Weinberg [1980] assumed that no critical attachments exist between the tongue and the pharyngoesophageal segment (a condition which may vary as a function of surgical method [Simpson et al., 1972]) through which lingual positioning could exert influence over the vibratory characteristics of the segment. They speculated that the higher lingual position with inherent higher impedance to air flow may mediate vowel F_0 through resultant increases in 'subglottic' pressure and the necessary force to overcome it. Moon and Weinberg [1987] suggested that a myoelastic mechanism may mediate vibration of the pharyngoesophageal segment, but did not identify the mechanism.

Laryngectomies are far from homgeneous with respect to the pharyngoesophageal segment as the vibratory body. The pharyngoesophageal segment represents an amalgamation of tissue remnant to the surgical procedure and varies anatomically and physiologically across individual patients [Weinberg, 1980]. Alteration of the 'supraglottic' structures is also subject to variation [Diedrich and Youngstrom, 1966] as a result of such factors as the advance of the carcinoma, the extent of the surgery, and the methods employed by the surgeons. Given this degree of diversity in both the source and the filter in esophageal speech, it would not be surprising to find that individual esophageal speakers may produce some degree of idiosyncratic variation in the acoustic characteristics associated with particular perceptual dimensions. This may mean that the perceptual saliency of F_0 reported for the present speaker may arise from the speaker's postsurgical capabilities, anatomically and physiologically. Future scaling studies with esophageal speech will have to determine whether this result is commonly obtained (and thus has perceptual salience) or whether it is simply an idiosyncratic and nonreplicable aspect of the present study.

The third perceptual dimension (rhotacization) correlates most highly with the F_3 measures. This result is consistent with Singh and Woods [1971], although other MDS studies [such as Fox, 1982, 1983, 1985a; Rakerd and Verbrugge, 1985] have failed to obtain a rhotacization dimension. However, the failure to obtain such a dimension in these studies almost certainly stems from the lack of a rhotacized vowel (e.g., [ɚ]) in the stimulus sets.

In summary, it seems that the basic perceptual structure of esophageal vowels is the same as that of laryngeal vowels. In many ways this is not a surprising result since one expects (in terms of the acoustic theory of speech production) [Fant, 1960] formant structure (and its perceptual correlate, vowel quality) to be primarily a function of the acoustic filter (i.e., articulatory tract configuration). We thus have no data which suggest that listeners need to modify their basic perceptual processes to identify esophageal vowels once they have extracted the necessary acoustic information (such as formant structure) from the speech wave. The unexpectedly high correlations between F_0 and dimension 2 will remain suspect until additional scaling studies have been conducted using different speakers.

The pattern of significant correlations between dimension 2 and the F_1-F_0 bark differences predicted by Syrdal and Gopal [1986] for laryngeal speech suggests that the F_1-F_0 bark difference is important in the discrimination of vowel height even when F_0 is very low. It is important to note that these results would have been eliminated if we had used Traunmüller's [1981] end correction, which may give evidence against its use.

The present study has described the perceptual structure of vowels produced by a single proficient esophageal speaker and the structure was found not to differ greatly from those of studies using individual laryngeal speakers. However, since individual esophageal speakers may differ in terms of proficiency and/or physiological structure, further research should attempt to determine the range of perceptual dimensions and their acoustical correlates found not only across different esophageal speakers, but across speakers using various forms of alaryngeal voice as well. We have raised the issue that some variation in perceptual dimensions may be found among different esophageal speakers. In addition, esophageal speech is not the only form of voice restoration for laryngectomized persons and different types of voice restoration may give further insight into the perceptual salience of the F_0 and F_1-F_0 measures. For example, speakers effectively using artificial larynges have limited or no capability for systematic alteration of F_0. Speakers using other forms of surgical voice restoration [e.g. Singer-Blom procedure; Singer and Blom, 1980] produce speech which is acoustically distinct from esophageal speech [Robbins, 1984] and which probably employs a different mechanism for regulating F_0 [Moon and Weinberg, 1987]. How the results from the present study pertain to the perceptual structure of vowels produced by these alaryngeal, nonesophageal speakers is unclear. Work is continuing in our laboratory using perceptual scaling methods using a wider variety of speakers and alaryngeal voice sources to address these issues.

Acknowledgments

The authors would like to thank Terrance Nearey, Jackson Gandour, and John Ohala for their comments on an earlier version of this paper. Versions of this paper were presented at the 1987 Spring Meeting of the Acoustical Society of America, Indianapolis, Ind., and at the 1987 Convention of the American Speech-Language-Hearing Association of America, New Orleans, La.

References

Bladon, A.; Henton, C.; Pickering, J.: Towards an auditory theory of speaker normalization. Lang. Comm. *4:* 59–69 (1984).

Carroll, J.D.; Chang, J.J.: Analysis of individual differences in multidimensional scaling via an *n*-way generalization of 'Eckart-Young' decomposition. Psychometrika *35:* 283–319 (1970).

Diedrich, W.; Youngstrom, K.: Alaryngeal speech (Thomas, Springfield 1966).

Doyle, P.C.; Danhauer, J.L.: Perceptual characteristics of esophageal and tracheoesophageal talkers. Annu. Convention of ASHA, Detroit 1986.

Fairbanks, G.: A voice and articulation drillbook (Harper, New York 1960).

Fant, G.: Acoustic theory of speech production (Mouton, s'-Gravenhage 1960).

Fox, R.A.: Individual variation in the perception of vowels: implications for a perception-production link. Phonetica *39:* 1–22 (1982).

Fox, R.A.: Perceptual structure of monophthongs and diphthongs in English. Lang. Speech *26:* 21–60 (1983).

Fox, R.A.: Multidimensional scaling and perceptual features: evidence of stimulus processing or memory prototypes? J. Phonet. *13:* 205–217 (1985a).

Fox, R.A.: Auditory contrast and speaker quality variation in vowel perception. J. acoust. Soc. Am. *77:* 1552–1559 (1985b).

Fujisaki, H.; Kawashima, T.: The roles of pitch and higher formants in the perception of vowels. IEEE Trans. Audio Electroacoust. *AU-16:* 73–77 (1968).

Gandour, J.; Harshman, R.: Crosslanguage differences in tone perception: a multidimensional scaling investigation. Lang. Speech *21:* 1–33 (1978).

Gandour, J.; Weinberg, B.: On the relationship between vowel height and fundamental frequency: evidence from esophageal speech. Phonetica *37:* 344–354 (1980).

Goldstein, L: Three studies in speech perception: features, relative salience, and bias. UCLA Working Papers Phonet. *39:* 1–87 (1971).

House, A.S.; Fairbanks, J.: The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. acoust. Soc. Am. *25:* 105–113 (1953).

Kruskal, J.B.; Wish, M.: Multidimensional scaling (Sage Press, Beverly Hills 1978).

Ladefoged, P.: A phonetic study of West African languages (Cambridge University Press, Cambridge 1968).

Ladefoged, P.: A course in phonetics; 2nd ed. (Harcourt Brace Jovanovich, New York 1982).

Lehiste, L.: Suprasegmentals (MIT Press, Cambridge 1970).

Miller, R.L.: Audiology tests with synthetic speech. J. acoust. Soc. Am. *25:* 114–121 (1953).

Moon, J.B.; Weinberg, B.: Aerodynamic and myoelastic contributions to tracheoesophageal voice production. J. Speech Hear. Res. *30:* 387–395 (1987).

Nearey, T.: Static, dynamic and relational properties in vowel perception. J. acoust. Soc. Am. (submitted, 1988).

Peterson, G.; Barney, H. Control methods used in a study of the vowels. J. acoust. Soc. Am. *24:* 175–184 (1952).

Pols, L.C.W.; van der Kamp, L.J.Th.; Plomp, R.: Perceptual and physical space of vowel sounds. J. acoust. Soc. Am. *46:* 458–467 (1969).

Rakerd, B.: Vowels in consonantal context are perceived more linguistically than are isolated vowels: evidence from an individual differences scaling study. Perception Psychophysics *35:* 123–136 (1984).

Rakerd, B.; Verbrugge, R.R.: Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels. J. acoust. Soc. Am. *77:* 296–301 (1985).

Robbins, J.: Acoustic differentiation of laryngeal, esophageal, and tracheoesophageal speech. J. Speech Hear. Res. *27:* 577–585 (1984).

Shepard, R.: Psychological representation of speech sounds; in David, Denes, Human communication: a unified view, pp. 67–113 (McGraw-Hill, New York 1972).

Simpson, I.C.: Smith, J.C.; Gordon, M.T.: Laryngectomy: the influence of muscle reconstruction on the mechanisms of oesophageal voice production. J. Lar. Otol. *86:* 961–990 (1972).

Singer, M.I.; Blom, E.C.: An endoscopic technique for restoration of voice after laryngectomy. Ann. Otol. Rhinol. Lar. *89:* 529–533 (1980).

Singh, S.; Woods, G.: Perceptual structure of 12 American English vowels. J. acoust. Soc. Am. *49:* 1861–1866 (1971).

Singh, S.; Woods, G.; Becker, G.M.: Perceptual structure of 22 prevocalic English consonants. J. acoust. Soc. Am. *52:* 1698–1713 (1972).

Sisty, N.; Weinberg, B.: Formant frequency characteristics of esophageal speech. J. Speech Hear. Res. *15:* 439–448 (1972).

Slawson, A.W.: Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. J. acoust. Soc. Am. *43:* 87–101 (1968).

Syrdal, A.K.: Aspects of a model of the auditory representation of American English vowels, Speech Commun. *4:* 121–135 (1985).

Syrdal, A.K.; Gopal, H.S.: A perceptual model of vowel recognition based on the auditory representation of American English vowels. J. acoust. Soc. Am. *79:* 1086–1100 (1986).

Terbeek, D.: A cross-language multidimensional scaling study of vowel perception. UCLA Working Papers Phonet. *37:* 1–271 (1977).

Terbeek, D.: Harshman, R.: Cross-language differences in perception of natural vowel sounds. UCLA Working Papers Phonet. *19:* 26–38 (1971).

Thurstone, R.; Chave, E.: The measurement of attitude: a psycho-physical method and some experiments with a scale for measuring attitude toward the church (University of Chicago Press, Chicago 1929).

Traunmüller, H.: Perceptual dimension of openness in vowels. J. acoust. Soc. Am. *69:* 1465–1475 (1981).

Trudeau, M.: A comparison of the speech acceptability of good and excellent esophageal and tracheoesophageal speakers. J. Commun. Dis. *20:* 111–119 (1987).

Weinberg, B.: Readings in speech following total laryngectomy (University-Park Press, Baltimore 1980).

Zwicker, E.; Terhardt, E.: Analytical expressions for critical-band rate and critical bandwith as a function of frequency. J. acoust. Soc. Am. *68:* 1523–1525 (1980).

Received: October 26, 1987
Accepted: June 23, 1988

Robert Allen Fox, Ph.D.
Speech and Hearing Science
Ohio State University
324 Derby Hall, 154 N Oval Mall
Columbus, OH 43210–1372 (USA)