Modeling perceived vowel height, advancement, and rounding

Patti Adank*, **Roel Smits**†, and **Roeland van Hout***

*University of Nijmegen, Nijmegen, The Netherlands
p.adank@let.kun.nl

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands†.

ABSTRACT

We investigated whether individual formant frequencies or distances between spectrally adjacent formant frequencies are more suitable for predicting perceived vowel articulation. The relationship between perceptual and acoustic representations of a set of vowel tokens was modeled. The acoustic representation consisted of measurements of F_0 and the first three formants. The perceptual representation consisted of coordinates representing each vowel token's perceived height, advancement or rounding. The comparison between the acoustic and perceptual representations was carried out through linear regression analysis. We concluded that formant frequencies were more suitable for modeling perceived articulation when the speaker-related variation was eliminated.

1 INTRODUCTION

Articulatory characteristics of vowels are generally described in terms of tongue height, tongue advancement, and lip rounding. It is generally assumed that articulatory vowel height correlates with F_1, articulatory tongue advancement correlates with F_2, while the relationship for articulatory rounding is less clear (cf. [1]). For the relationship between *perceived* articulatory characteristics and acoustics, different solutions have been suggested. On one hand, [2], [3] and [4] propose predicting perceived height using the distance between F_1 and F_0. [3] predicts that perceived advancement can be modeled using the spectral distance between F_3 and F_2. Predictions made in [2] and [4] are based on perception experiments using synthetic speech and [3] observes this pattern in acoustic vowel data. On the other hand, [5] proposes to model perceived height using F_1 and perceived advancement using F_2. [5] compares acoustic measurements and perceptual judgments on the same set of vowel data. There are two discrepancies between results reported in [5] and [2], [3], and [4] regarding the perception of height and advancement. First, [2], [3] and [4] propose that distances between spectrally adjacent formant frequencies are more suitable for modeling perceived articulation because distances are relatively constant across speakers and therefore incorporate some sort of speaker normalization. However, [5], in speech is used from 10 male and 10 female speakers, reports that individual formant frequencies may be suitable after all. Second, [5] finds that F_0 contribute only marginally to the model for height and that F_3 does not contribute to the model for advancement, while [2], [3], and [4] predict that F_0 is important for the perception of height and [3] predicts that F_3 is important for advancement. The present research focuses on the two discrepancies between [5] and [2], [3], and [4]. The first discrepancy is addressed as follows. [5] does not explicitly compare the models of perceived height and advancement, and rounding for individual formant frequencies with the models for formant frequency distances. To establish whether individual formant frequencies or formant frequency distances are more suitable for modeling perceived articulation of height, advancement, and rounding, the models for individual formant frequencies were compared to the models for formant frequency distances. The second discrepancy is addressed as follows. F_0 and F_3 may improve the fit of perceptual models of height and advancement when expressed in relation to F_1 and F_3, respectively, through the use of F_1-F_0 and F_3-F_2. However, the finding that F_0 and F_3 do not, or only marginally, contribute to the perceptual models may be due to the fact that listeners in [5] had to perform speaker normalization. The listeners had to judge read speech from 20 speakers (male and female), whereas in [2], [4], synthetic speech was used that was made to sound as if produced by one speaker. The second discrepancy was thus investigated through the use of acoustic speaker normalization, as done in [5].

2 METHOD

2.1 Perceptual representation

A perceptual representation of a set of vowel tokens - judgments of vowel tokens' height, advancement, and

ISBN 1-876346-48-5

rounding - was obtained through a listening experiment with seven phonetically trained listeners. The listeners had to be phonetically trained seeing as they had to be familiar with the IPA vowel quadrilateral. [5] and [6] describe this experiment in detail. The stimuli were read vowels in a /sVs/-context, where "V" stands for one of the nine monophthongal vowels of Dutch, /ɑ/, /a/, /ɛ/, /ɪ/, /i/, /ɔ/, /u/, /ʏ/, /y/, produced by 10 female and 10 male speakers of Dutch. For each stimulus, the coordinates corresponding to the chosen location in the quadrilateral and rounding scale represented perceived height, advancement, and rounding. The listeners were required to judge each vowel token's perceived height, advancement, and rounding. Height and advancement judgments were obtained by asking the subjects to locate each stimulus at the appropriate location in a response area shaped like the IPA 1996 vowel quadrilateral. The horizontal axis of the quadrilateral represented tongue advancement (from left to right representing front to back). The vertical axis represented vowel height (from bottom to top representing low to high perceived vowel height). Rounding was judged in a separate rounding scale, with the left side representing more lip spreading and the right side representing more lip rounding. The coordinates corresponding to the locations in the quadrilateral (height and advancement), and the rounding scale (rounding) served as perceptual descriptions of each of the 180 vowel tokens. The mean values across the seven listeners were used as the perceptual representation of each of the 180 vowel tokens.

2.2 Acoustic representation

The acoustic representation was made using the same 180 vowel tokens used for the perceptual representation. The acoustic measurements were obtained automatically using a program for formant measurement and formant tracking described in [7]. For each vowel token, the formant tracks were visually inspected and adjusted by hand whenever this was thought necessary, using the built-in interface of the program. The measurements are discussed in more detail in [6].

3 COMPARISON OF PERCEPTUAL AND ACOUSTIC REPRESENTATIONS

The comparisons between the perceptual and acoustic representations of the experiment and the (normalized) acoustic data were performed using linear regression analysis (LRA). Several combinations of the acoustic predictor variables were regressed onto the three perceptual criterion variables height, advancement, and rounding. A total of 10 combinations of predictors was tested. Seven LRAs were carried out with a single predictor variable (F_0, F_1, F_2, F_3, F_1-F_0, F_2-F_1, and F_3-F_2) and three with two predictors (F_0, F_1 and F_1, F_2 and F_2, F_3). The results for the distances (e.g., F_1-F_0) were compared to the single predictors (e.g., F_1) and to the pairs of predictors (F_0, F_1).

Table 1: Percentages explained variance ($R^2 \times 100\%$) for combinations of F_0, F_1, F_2, and F_3 in Hz for the linear regression analyses for Height ("H"), Advancement ("A"), and Rounding ("R"). Only significant results are shown, $p<0.001$.

$R^2 \times 100$		H	A	R
Individual	F_0	45	-	-
	F_1	72	10	13
	F_2	12	80	39
	F_3	-	-	-
Pair	F_0, F_1	76	10	13
	F_1, F_2	79	85	60
	F_2, F_3	19	85	41
Distance	F_1-F_0	76	11	15
	F_2-F_1	36	84	20
	F_3-F_2	19	75	18

Table 1 shows $R^2 \times 100\%$ for each LRA. For height, the percentage for the formant frequency distance F_1-F_0 (76%) is higher than for individual F_0 (45%), individual F_1 (72%), and equally high as the pair F_0, F_1 (76%). However, the score for F_1-F_0 is lower than for the pair F_1, F_2 (79%). This implies that perceived height could be modeled better through the use of a pair of predictor variables: F_1, F_2, than through the distance F_1-F_0. For advancement, the pattern is different: the percentages for the formant frequency distance F_2-F_1 (84%) is higher than for individual F_2 (80%) and for individual F_1 (15%), but lower than the pair F_1, F_2 (85%). In addition, F_3-F_2 (75%), suggested in [3] for modeling perceived advancement does worse than F_2 (80%) and F_2, F_3 (85%). For rounding, it can be observed that individual F_2 (39%) performs better than the distances F_3-F_2 (18%) F_2-F_1 (20%), but worse than the pair F_1, F_2 (60%) and the pair F_2, F_3 (41%).

Table 1 is equivocal on whether individual formant frequencies or formant frequency distances are more suitable for modeling perceived articulation; height could be modeled best using the formant frequency distance, while advancement and rounding could be modeled best using a pair of predictors. A possible explanation for these results is that the acoustic measurements show considerable variation depending on the anatomical/physiological characteristics of speakers (e.g., differences in the larynx and vocal-tract of male and female speakers). This anatomical/physiological variation may have affected (or even contaminated) the results of the models. The anatomical/physiological variation can be minimized using a procedure for speaker normalization. In [5], a transformation to z-scores per speaker, proposed in [8], was found to be most effective at minimizing anatomical/physiological varia-

ISBN 1-876346-48-5

tion in accordance with the judgments of phonetically trained listeners.

The raw data in Hz (F_0, F_1, F_2, and F_3) were transformed to z-scores using [8] to minimize the anatomical/physiological variation in the acoustic data. Two multivariate analyses of variance were carried out to establish the relative proportion of variance in the acoustic representation arising from the variation sources gender and vowel. An additional analysis was carried out on the perceptual representation. The portion of explained variance was expressed using η^2.

Table 2: Results for the multivariate analyses of variance: η^2 for each significant factor, for the raw and normalized data, and for the perceptual representation: Height ("H"), Advancement ("A"), and Rounding ("R") ($p<0.001$).

η^2	Vowel	Gender	Vowel×Gender
F_0	0.142	0.683	-
F_1	0.902	0.299	0.325
F_2	0.919	0.372	-
F_3	0.329	0.391	-
zF_0	0.327	-	-
zF_1	0.944	-	-
zF_2	0.970	-	-
zF_3	0.524	-	-
H	0.990	-	-
A	0.986	-	-
R	0.968	-	-

Table 2 shows, for the acoustic representations, that most of the anatomical/physiological variation present in the raw data in Hz was minimized after the transformation to z-scores. The proportion of the variation that could be attributed to gender is relatively high for raw F_0, F_1, F_2, and F_3, while this proportion was not significantly different from 0 for normalized zF_0, zF_1, zF_2, and zF_3. In addition, the proportion of the variation that could be attributed to the vowel is higher for normalized zF_0, zF_1, zF_2, and zF_3 than for raw F_0, F_1, F_2, and F_3. Table 2 shows similarities in the results of the normalized acoustic data and the perceptual representation, both show high proportions vowel-related variance, and no significant proportions for gender, nor for the interaction between gender and vowel. It appears that the judgments of the phonetically trained listeners do not vary systematically depending on the speaker's gender. Summarizing, the results show that the normalized acoustic representation resembles the perceptual representation more than the raw acoustic representation, it shows considerably less anatomical/physiological variation than the raw acoustic representation, and it displays a larger proportion of vowel-related variation than is the case for the raw acoustic representation.

To establish if the LRAs displayed in Table 1 were affected by the anatomical/physiological variation present in the raw acoustic representation, the same set of LRAs were run again, this time with the normalized acoustic variables.

Table 3: Percentages explained variance ($R^2 \times 100\%$) for combinations of zF_0, zF_1, zF_1, and zF_1 (transformed to z-scores) for the linear regression analyses for Height ("H"), Advancement ("A"), and Rounding ("R"), $p<0.001$.

$R^2 \times 100$		H	A	R
Individual	zF_0	24	-	8
	zF_1	82	11	15
	zF_2	14	86	40
	zF_3	-	-	14
Pair	zF_0, zF_1	82	11	16
	zF_1, zF_2	87	90	65
	zF_2, zF_3	22	86	52
Distance	zF_1-zF_0	67	8	13
	zF_2-zF_1	70	67	3
	zF_3-zF_2	21	56	4

Table 3 shows the results for the LRAs on the z-scores. It can be seen that the individual formant frequencies zF_1 (82%), zF_2 (86%), and zF_2 (40%) show higher scores for height, advancement, and rounding, respectively, than is the case for the formant frequency distances: zF_1-zF_0 (67% for height), zF_2-zF_1 (67% for advancement), zF_3-zF_2 (56% for advancement), zF_2-zF_1 (3% for rounding), zF_3-zF_2 (4% for rounding). The same pattern can be observed for the pairs of predictors: all pairs show higher scores than the distances (e.g., zF_1, zF_2 is 90% while zF_3-zF_2 is 56%). Given the results in Table 3, it seems justified to assume that individual formant frequencies (and pairs of spectrally adjacent formant frequencies) are more suitable for modeling perceived height, advancement, and rounding, when the variation in the acoustic data due to the speaker's anatomical/physiological characteristics is minimized.

A remarkable observation from Table 3 is that individual zF_0 contributes 24% to the model for height, individual zF_1 contributes 82%, but the pair zF_0, zF_1 also contribute 82%. zF_0 did not contribute to the model for height when zF_1 was also added. In addition, the distance zF_1-zF_0 (67%) performs much worse than zF_1 (82%). This pattern in the results is not found for the combinations of raw F_0 en F_1 in Table 1. The correlation between the two variables was calculated, to establish if the difference in the results for raw and normalized F_0 and F_1 was different before and after normalization. Before normalization, no significant correlation was found between F_0 and F_1, while after normalization, a correlation score of -0.46 was found (Pearson's r). This indicates that transforming raw data to z-scores causes zF_0 to correlate negatively with zF_1. The high correlation score between normalized zF_1 and zF_0 can be interpreted as

ISBN 1-876346-48-5

that both contain considerable information about the vowel's identity, which may have been obscured in raw F_0 by the presence of gender-related variation (cf. Table 2). It seems plausible that, when all gender-related variation is eliminated from the normalized vowel data, the remaining variation in F_0 is of a phonemic nature (i.e., vowel intrinsic pitch). This phonemic information in zF_0 is probably also present in zF_1. This redundancy was probably the reason why entering zF_0 as well as zF_1 to the model for height did not result in an improvement compared to entering only zF_1.

4 DISCUSSION

The results show that perceived height could be modeled better using zF_1, than with zF_1-zF_0, and that perceived advancement could be modeled better using zF_2 than using zF_3-zF_2, or zF_2-zF_1. Perceived height and advancement were thus modeled best using individual formant frequencies, as predicted in [5], and not using distances between spectrally adjacent formant frequencies, as predicted in [2], [3], and [4], but only after applying speaker normalization. Furthermore, because the listeners were found not to be affected by anatomical/physiological differences between speakers and the same was found for the acoustic data normalized following [8], it can be hypothesized that the predictions about formant frequency distances as made in [2] and [4] cannot be generalized to a task in which listeners have to perform speaker normalization (as was the case in the present study and in [5]). Note that [2] and [4] use synthetic speech material and that [5] use speech from male and female speakers.

The difference in performance between the individual normalized formant frequencies and the normalized formant frequency differences may be explained if it is assumed that both types of representations make different assumptions about the process of vowel perception. A representation such as F_3-F_2, using the terminology expressed in [9], is of a *vowel-intrinsic* nature whereas zF_2 is of a *vowel-extrinsic* nature . A transformation of formant values using speaker normalization can be regarded as a way of obtaining a more perceptually relevant representation. Intrinsic and extrinsic representations predict that normalization occurs at different moments in vowel processing. If vowel normalization is an intrinsic process, then normalization is essentially some sort of efficient preprocessing of the acoustic stimulus occurring at a peripheral auditory stage, before more central processing takes place. Intrinsic processing requires only one vowel, whereas extrinsic processing requires information across several, if not all, vowels per speaker. This implies that information about those vowels should be retained in memory until all vowels for a speaker are heard. However, it seems unlikely that this type of information is retained at a peripheral auditory stage, it seems more likely that it is retained during more central processing. However, for vowel-extrinsic processing to be successful, the listener must learn the speaker's vowel system. Listeners may do this by making a temporary prediction that is adjusted whenever new information arrives. The results of the present paper for F_0 may be interpreted as that listeners use the fundamental frequency of a vowel as a cue to learn the speaker's vowel system. However, these hypotheses can neither be confirmed nor rejected due to the correlational nature (i.e. the use of LRA) of the present study; they must be investigated further before more conclusive remarks can be made.

ACKNOWLEDGMENTS

This research is supported by the Stichting Spraaktechnologie.

REFERENCES

[1] K. N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, 1998.

[2] H. Traunmüller, "Perceptual dimensions of openness in vowel," *Journal of the Acoustical Society of America*, vol. 69, pp. 1465–1475, 1981.

[3] A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *Journal of the Acoustical Society of America*, vol. 79, pp. 1086–1100, 1986.

[4] R. P. Fahey, R. L. Diehl, and H. Traunmüller, "Perception of back vowels: Effects of varying $F_1 - F_0$ Bark distance," *Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2350–2357, 1995.

[5] P. Adank, R. van Hout, and R. Smits, "A comparison between human vowel normalization strategies and acoustic vowel transformation techniques," in *Proc. of EUROSPEECH '01*, Aalborg, Denmark, 2001, pp. 481–484.

[6] P. Adank, *Vowel normalization: a perceptual-acoustic study of Dutch vowels*, Ph.D. thesis, University of Nijmegen, Nijmegen, The Netherlands, 2003.

[7] T. M. Nearey, P. F. Assmann, and J. M. Hillenbrand, "Evaluation of a strategy for automatic formant tracking," *Journal of the Acoustical Society of America*, vol. 112(5), pp. 2323, 2002.

[8] B. M. Lobanov, "Classification of Russian vowels spoken by different speakers," *Journal of the Acoustical Society of America*, vol. 49, pp. 606–608, 1971.

[9] T. M. Nearey, "Static, dynamic, and relational properties in speech perception," *Journal of the Acoustical Society of America*, vol. 85, pp. 2088–2113, 1989.