SIMILARITY BETWEEN SCHWA-LIKE STIMULI MANIPULATED IN TERMS OF MALE, FEMALE AND CHILD CHARACTERISTICS

Louis C.W. Pols

1. INTRODUCTION

Stationary vowel-like signals can be described by two main signal aspects. The one is fundamental frequency which can vary between, say 80 and 500 Hz, corresponding to the one-dimensional perceptual attribute (voice) pitch going from low to high. The other signal aspect is the frequency spectrum (described in terms of harmonic levels, bandfilter values, linear prediction coefficients, formant frequencies, or whatever), with which corresponds the multidimensional perceptual attribute timbre or vowel quality.
By using technical means one can, in principle, generate any spectral form with any fundamental frequency. However, in natural speech there are many practical limitations; this reduces the possible set of speech signals to a restricted subset. Furthermore, natural changes in one parameter, for instance a pitch shift from a male to a female voice, are not independent from changes in the other (spectral) parameters. Slawson (1968) showed that a doubling of the fundamental frequency needs a "compensation" in terms of a 10% higher value for the formant frequencies in order to keep the same vowel quality. Plomp (1976) gives an overview of the psychophysical literature and the speech literature about the interaction between pitch and timbre. Despite the fact that one can speak, and certainly sing, one and the same vowel with a wide range of fundamental frequencies, there seem to be default relationships between fundamental frequency and formant values for a specific vowel from a specific speaker. It is rather probable that vowel quality is interpreted relative to its fundamental frequency, although this is only one of the aspects of speaker normalization (Pols, 1977). This interaction between fundamental frequency and spectral characteristics in a speech-like signal is the main topic of this paper.

2. METHOD

We started from three "standard signals": the vowel [œ] spoken by a man, a woman, and a child. In fact we used synthetic, stationary, 175-ms signals. The average values for fundamental frequency and formants were taken from data gathered by Weenink (1985) in a project on speaker normalization, including vowel data from 10 men, 10 women, and 10 children. The average fundamental frequencies for these male, female, and child speakers were 140, 240, and 310 Hz respectively. The average formant values for the vowel [œ] are given in Fig. 1. The three diagonal points represent the "standard" male, female, and child vowel [œ] generated with their "natural" fundamental frequency and formant values by using a vowel generating program based on digital filters and a periodic source. The six off-diagonal signals in Fig. 1 represent various "unnatural" combinations of the three fundamental frequency values and the three formant combinations. This resulted in signals which could be interpreted as "a male [œ] with a female pitch" (nr. 4), or "the [œ] from a child with a male

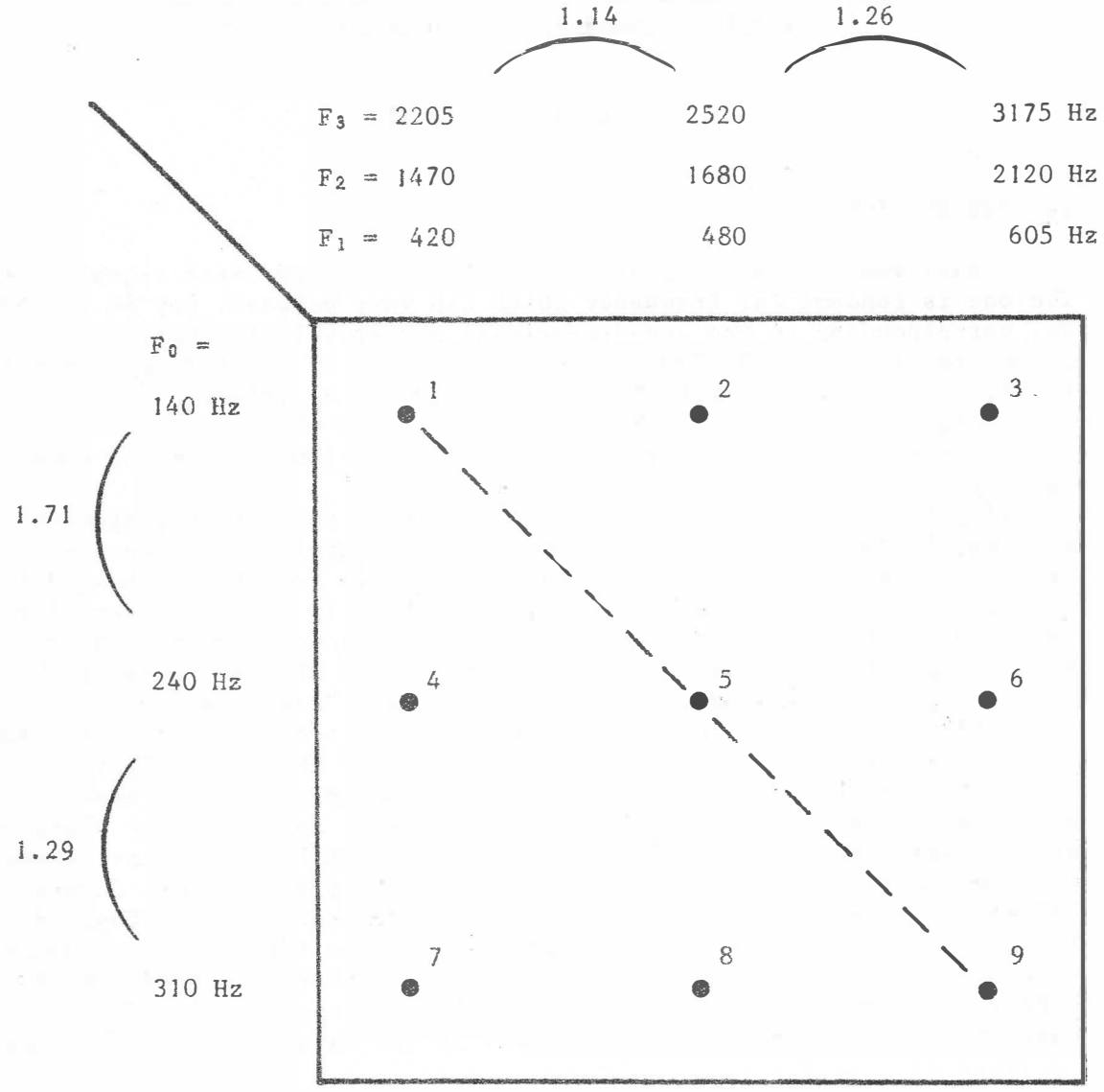

Fig. 1 Idealized representation of the nine stimuli in a square. The formant values and the fundamental frequency for every stimulus are indicated, as well as the ratios between these values.

pitch" (nr. 3). In fact the set of nine stimuli did not all sound like allophonic variations of /œ/ but some of them could also be labeled as /ɛ/ or /y/. None of the signals sounded very unnatural and they all evoked vowel-like reactions, also from naive listeners, especially when a second version of the stimuli was presented in which we replaced the stationary fundamental frequency by a downward shift over 10%.
These nine stimuli were presented to eight listeners in a triadic comparison experiment. The subjects got written instructions. In fact we ran two separate experiments: one with stationary fundamental frequencies resulting in a rather synthetic quality (Experiment 1), and one with a

downward-going fundamental frequency which resulted in somewhat more natural-sounding stimuli (Experiment 2).
In a triadic comparison experiment all possible combinations of three stimuli out of the total set of nine are presented to one listener at a time. By pairwise comparison of the three signals under investigation he/she has to judge which of the two stimuli are most similar, and after that which pair of stimuli of the two remaining pairs is most dissimilar. The subject can listen as often as he wants to any of the three stimuli by pressing the appropriate stimulus buttons on a triangular box. The judgment can be made by pressing one of the three response buttons positioned between the three stimulus buttons. Stimulus generation and response processing is computer controlled. A response can only be given if all three stimuli have at least been listened to once, and the most dissimilar response cannot be the same as the most similar response in one triad. The subject gets visual feedback as to which response has already been given. Every subject gets a different random sequence of all triads. The results of such an experiment are as a rule presented in terms of a dissimilarity matrix of which only the lower, or upper, half is filled. Every time a pair of stimuli is judged most dissimilar, the corresponding matrix value is increased with two points, a most similar judgment results in no extra points, and the remaining pair gets one point. The resulting dissimilarity matrix, in which the ordinal relations are relevant and the absolute relations are not necessarily, is very suitable as input for so-called non-metric multidimensional scaling programs of which Kruskal's (1964) MDSCAL is the best known. The output of such a program is in terms of a configuration of N points in n dimensions. The N points represent the N stimuli, and the Euclidean distances between these N points are rank-order correlated as well as possible with the input dissimilarity matrix. These programs use iterative optimization procedures to change the positions of the points until optimal match. The more dimensions are allowed the better this match can be. The most probable dimensionality of the solution is generally chosen on the basis of interpretability of the configuration, on an "elbow" in the curve representing the amount of stress as a function of the number of dimensions, or on the basis of deviations from random input (Monte Carlo procedures, e.g. Wagenaar and Padmos, 1971). By using a so-called Minkowski metric different from m=2, one can also use other distance models like city block (m=1).
Certain programs not just use the ordinal relations in the input data but also the absolute values. Originally only one individual or cumulative similarity matrix could be processed, more recent programs allow for input of all individual matrices, resulting in a stimulus configuration plus a so-called subject space. The best-known program of this last category is INDSCAL (Carroll and Chang, 1970).

3. EXPECTATIONS

If pitch and vowel quality changes were completely independent, and if differences between male, female, and child voices were of about equal magnitude, then a three by three square representation of the nine stimuli as in Fig. 1 would be the expected one. Changes from square to rectangular

(either lying or standing) would be related to higher emphasis on either pitch or vowel quality changes. Other changes from a rectangular form could be expected if for instance the three standard stimuli on the diagonal would be interpreted in a way substantially different from the other six stimuli. One could imagine that the standard male, female, and child vowel [œ] would be interpreted as almost identical because we are all well trained to ignore the speaker-specific variation and take full notice of the vowel similarity. This could result in a deviated rectangle from which the upper left point and the lower right point move inward to the centre. If all individual matrices were used as input to INDSCAL-type programs, one could imagine that the subject space would reflect the relative importance of pitch versus vowel quality for every individual listener. By stretching or shrinking one dimension relative to the other, individual deviations from the ideal three-by-three square would be possible. By combining the matrices of both experiments (the one with stationary pitch and the one with downward pitch), we might get some indication of the specific differences between both conditions.

4. RESULTS

The cumulative dissimilarity matrices, derived by summing the matrices of the eight subjects, are given in Table 1. The lower left part represents the data of Experiment 1 with stationary fundamental frequency, and the upper right part represents the data of Experiment 2 with downward pitch. Both matrices are processed by MINISSA, a non-metric multidimensional scaling program very similar to Kruskal's MDSCAL. This program was available to us via the MDS-package of the Technical Centre of the University of Amsterdam. Analyses were done in 4, 3, and 2 dimensions. It was very clear from the results that the most probable solution is two-dimensional, see Figs. 2 and 3. Since the best match of the point configuration with the input dissimilarity matrix is based on interpoint distances, the orientation of the axes is in principle irrelevant. For the solutions in Figs. 2 and 3 the axes are rotated in such a way that

Table 1 Cumulative dissimilarity matrix of Experiment 1 (lower left matrix) and Experiment 2 (upper right matrix).

	1	2	3	4	5	6	7	8	9
1	-	14	66	30	43	100	56	54	108
2	15	-	58	42	30	91	54	60	96
3	50	47	-	83	85	10	91	96	26
4	35	47	82	-	13	69	16	35	88
5	38	34	81	7	-	65	48	26	70
6	97	91	25	60	62	-	80	75	2
7	67	63	95	28	31	76	-	10	72
8	74	72	88	44	38	69	10	-	54
9	103	97	43	73	71	8	58	37	-

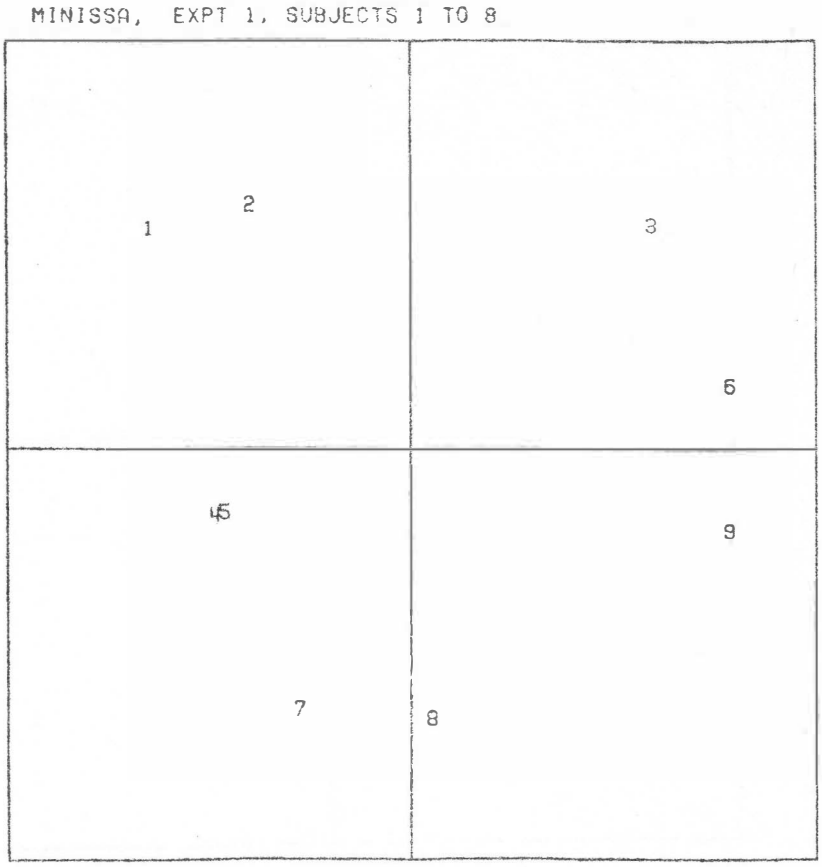

Fig. 2 Two-dimensional representation of the nine stimuli. In the upper left-hand corner the multi-dimensional scaling program, the number of the experiment, and the subject number, or the group of subjects for which this analysis is performed, are specified. The horizontal axis is dimension I and the vertical axis is dimension II. Since the coordinate values of the points are normalized, no further specifications of axis values is given.

dimension I explains most of the variance. Dimension I can be interpreted as the "formant dimension" and dimension II as the "pitch dimension". However, by comparing these two configurations with the ideal configuration of Fig. 1, it will immediately become clear that the distance between male and female formant values is much smaller than between female and child. The pitch steps between male, female, and child voice are about the same, although the total range for the high-formant stimuli (nr. 3, 6, and 9) is much smaller. The difference between stimuli 1 and 2, and between 7 and 8 is already rather small, but stimuli 4 and 5 are judged to be almost identical. These and other effects can partly be the result of summing the data over all eight subjects.

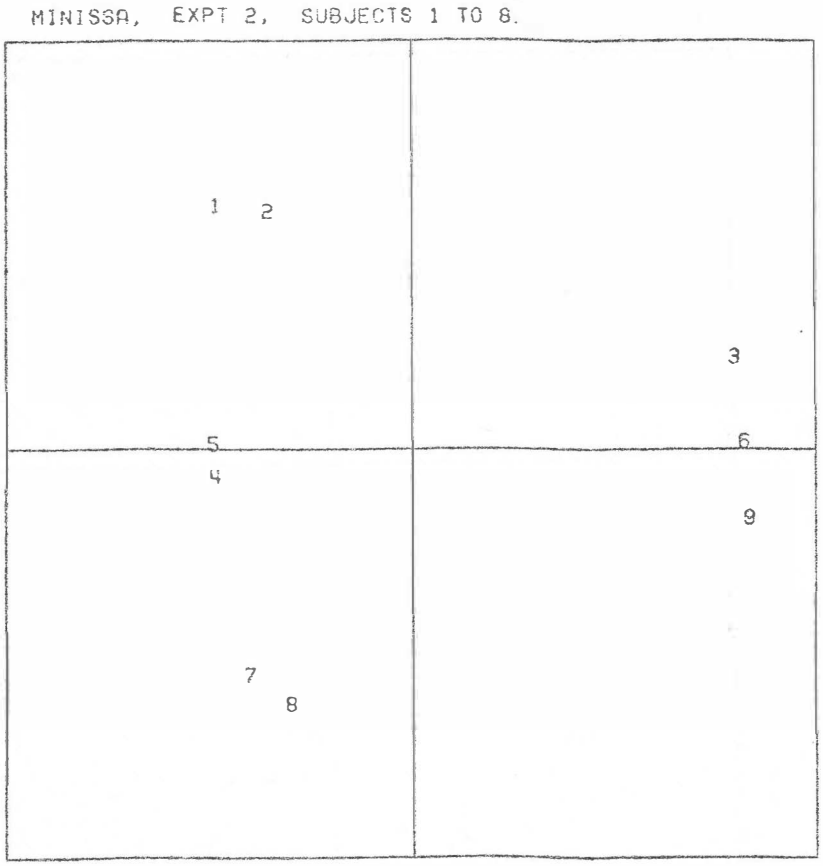

Fig. 3 Similar to Fig. 2.

Therefore we also analyzed the individual matrices with INDSCAL, resulting in an "object space" with the nine stimuli as points, and a "subject space" with the weights for the eight listeners. Figs. 4 and 5 give the solution in two dimensions for Experiment 1 and 2, respectively. In this solution one is no longer free to rotate the axes, so it is the more interesting that here again we find the same formant and pitch dimension. The stimulus configurations for both experiments are rather similar and the outer corners almost form a square. However, in this solution also stimuli representing male and female formant values (1, 4, and 7 versus 2, 5, and 8) are much closer together than the child stimuli (3, 6, and 9). There is also some indication that the female and child pitch stimuli (4, 5, and 6 versus 7, 8, and 9) are judged more similar than the female and male pitch stimuli (4, 5, and 6 versus 1, 2, and 3). In the object space of Experiment 1 we can see that all eight subjects fit rather well (none of them deviates very much from the unit circle, furthermore the amount of variance explained is 85.3%), although for subject 3 dimension II is relatively more important, and for subject 4 it is dimension I. A similar situation exists for Experiment 2, see Fig. 5. The amount of variance explained here is 86.1%. The more inward position for subject 4 is an indication that this subject's behaviour is not described very well with this configuration.

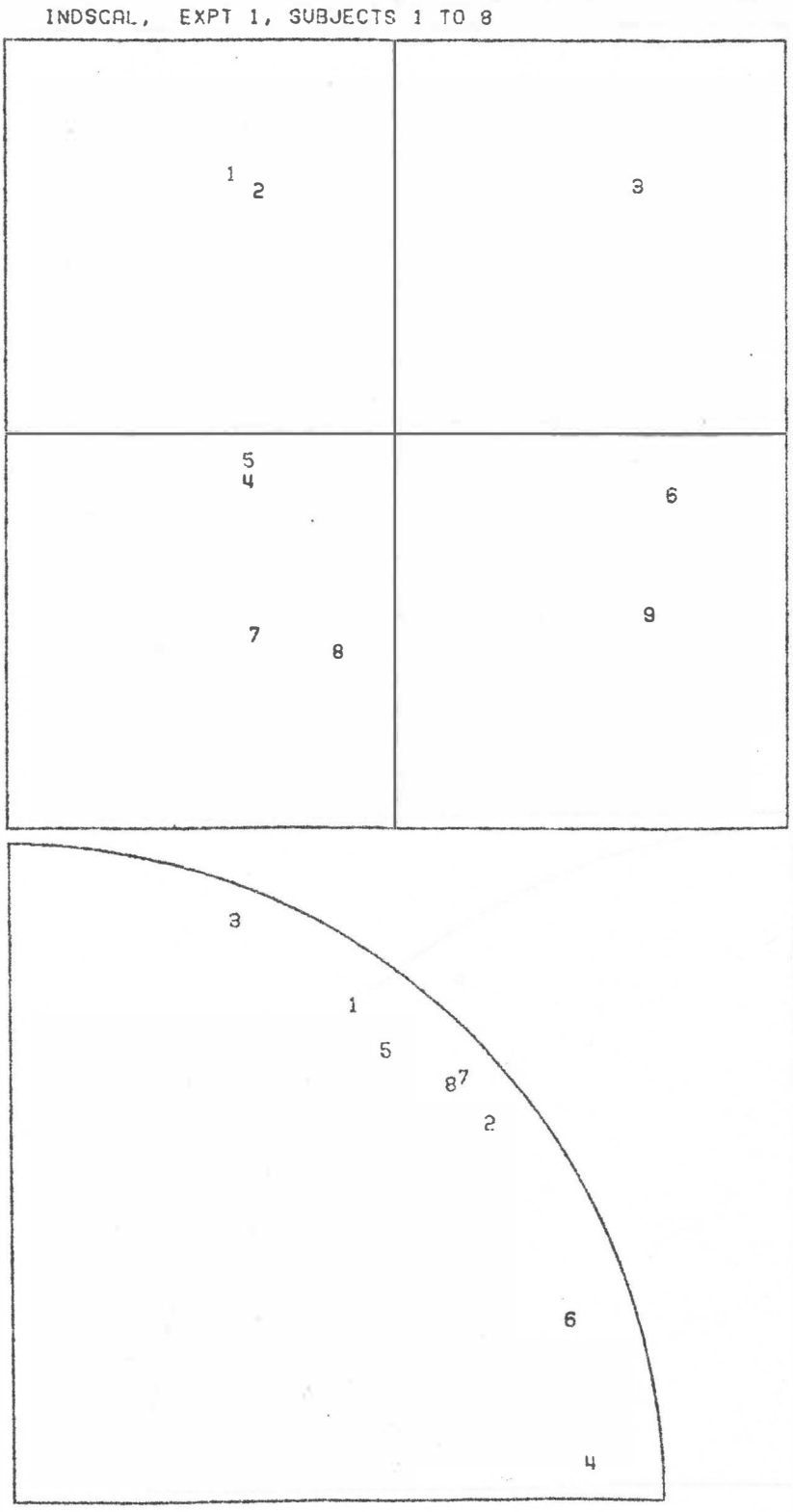

Fig. 4 Two-dimensional INDSCAL solution. The nine stimuli are represented in the object space in the upper part and the weights of the eight listeners are represented in the subject space in the lower part.

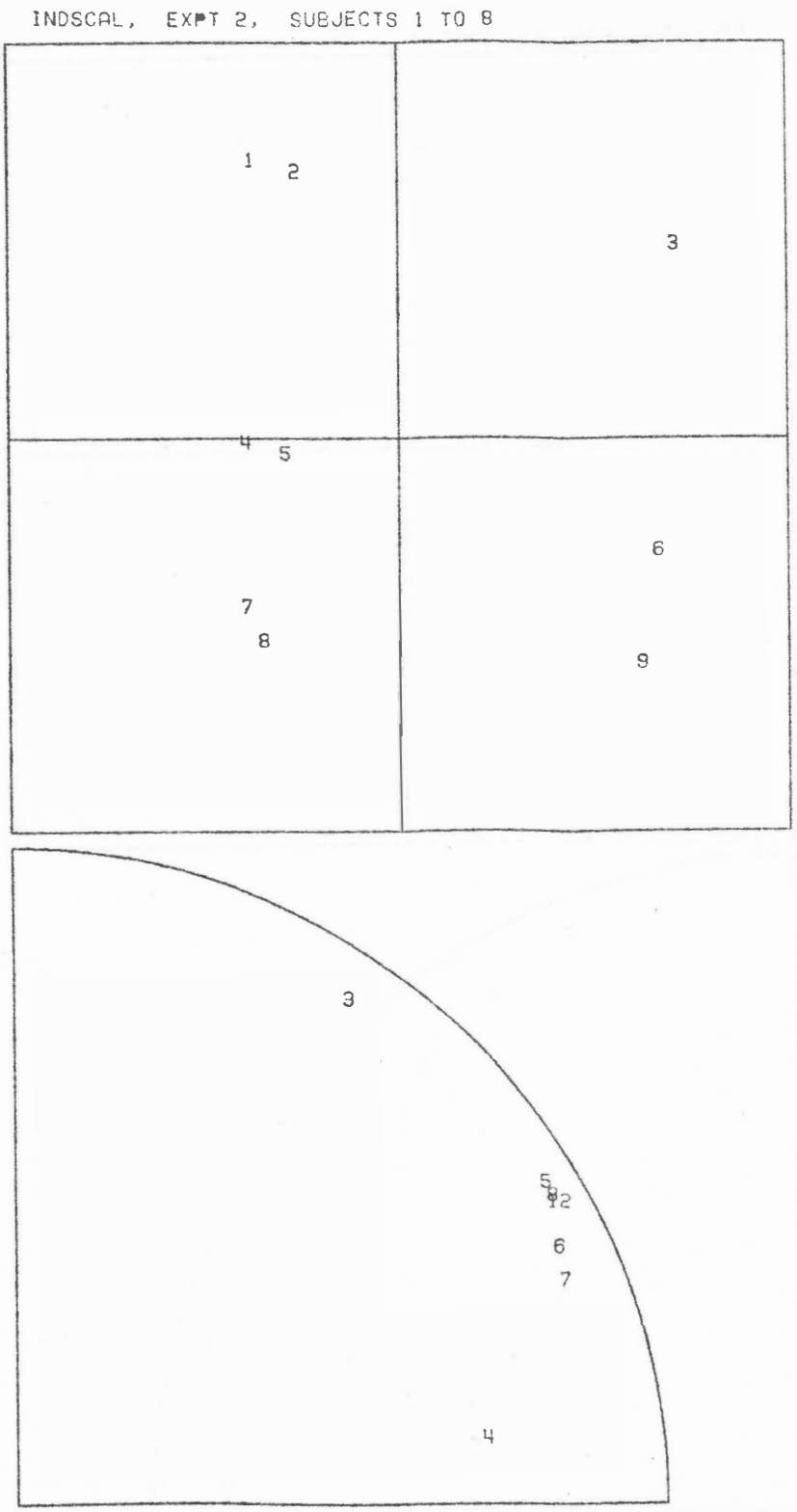

Fig. 5 Similar to Fig. 4, but this time for the data of Experiment 2.

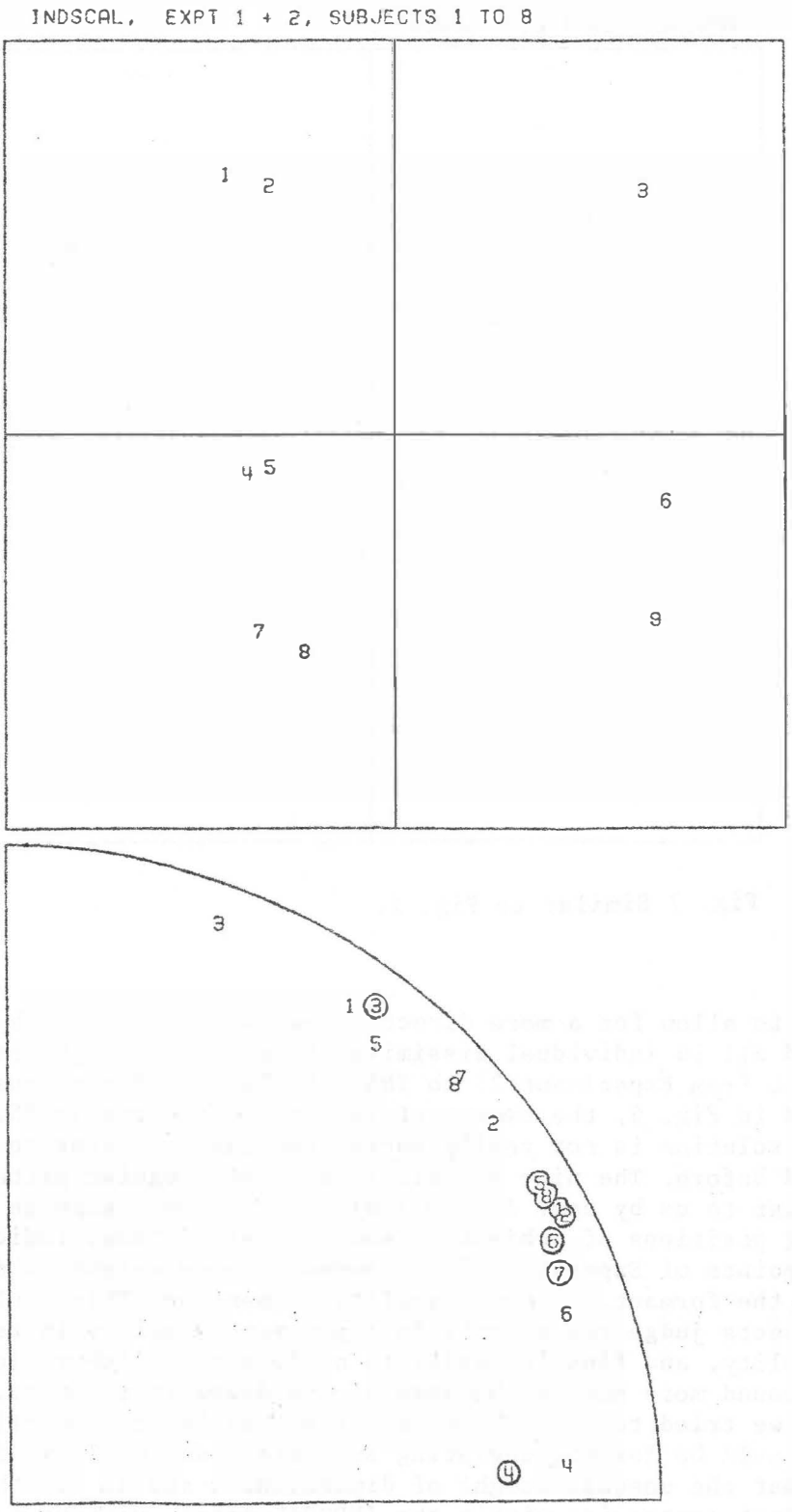

Fig. 6 Two-dimensional INDSCAL solution for the combined data of Experiment 1 and 2. The nine stimuli are represented in the object space in the upper part and the weights of the eight listeners are represented in the subject space in the lower part. The subject points from the second experiment are encircled.

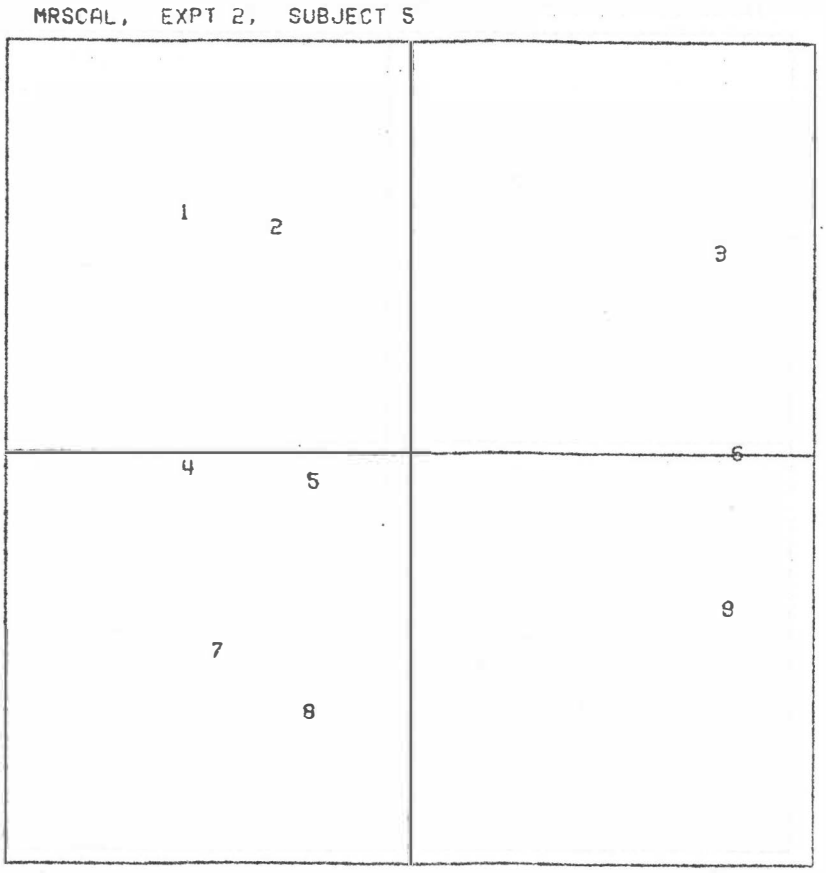

Fig. 7 Similar to Fig. 2.

In order to allow for a more direct comparison between both experiments we presented all 16 individual dissimilarity matrices (eight from Experiment 1 plus eight from Experiment 2) to INDSCAL. The two-dimensional solution is presented in Fig. 6, the amount of variance explained is 85.3%, so this combined solution is not really worse than the two separate solutions presented before. The nine stimuli form a very regular pattern which should be familiar to us by now. In the subject space we recognize again the deviating positions of subjects 3 and 4. There is some indication that all subject points of Experiment 2 put somewhat more weight on dimension II which is the formant, or vowel quality, dimension. This could mean that most subjects judge the stimuli in Experiment 2 mainly in terms of their vowel quality, and find it easier to neglect the pitch variations when the stimuli sound more natural because of the downward movement.
Finally, we tried to find out what the actual underlying stimulus configuration could be for the deviating subjects 3 and 4. There is always a chance that the unequal weight of dimensions I and II for these subjects was the best compromise within the INDSCAL-model, without really being a good description. Therefore we analyzed the individual matrices from Experiment 2 of subjects 3 and 4, and for comparison also that of an "average" subject nr. 5. Since we wanted to compare these solutions with

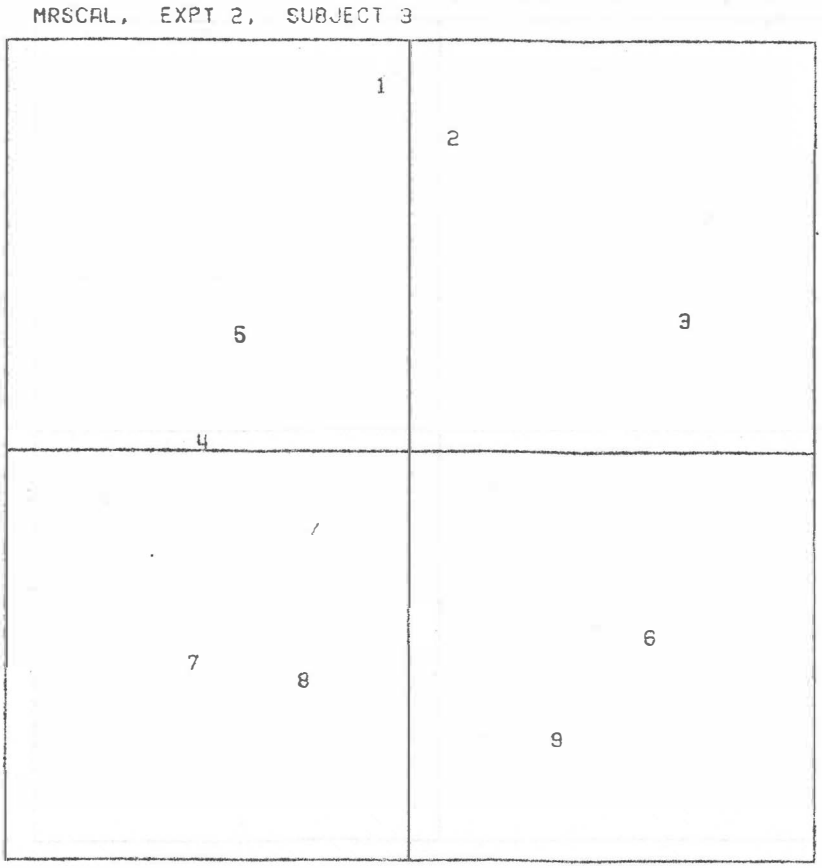

Fig. 8 Similar to Fig. 2.

INDSCAL's object space, and since INDSCAL is a metric procedure, we figured that the individual matrices also should be analyzed with a metric scaling program. In the aforementioned MDS-package, program MRSCAL was available for that purpose. For the average subject 5 we see indeed the common regular pattern, see Fig. 7. However, the solutions for subjects 3 and 4, see Figs. 8 and 9 respectively, deviate substantially from this pattern. For subject 3 the configuration seems to be rotated with somewhat deviated positions for stimuli 4 and 5, see Fig. 8. For subject 4 the stimulus configuration (see Fig. 9) is difficult to interpret: only stimuli 3, 6, and 9 are clearly separated from the other six stimuli, which probably explains why the weight for dimension I in Fig. 6 is still rather high. Dimension II of Fig. 9 has hardly any resemblance with dimension II in the overall space represented in Fig. 6, which is also reflected by the very low weight along dimension II in the object space for subject 4.

5. CONCLUSIONS

In an experimental situation like the one we used in our listening experiment (triadic comparison of nine synthetic vowel-like stimuli), one should not be too surprised to find that the results are more a reflection

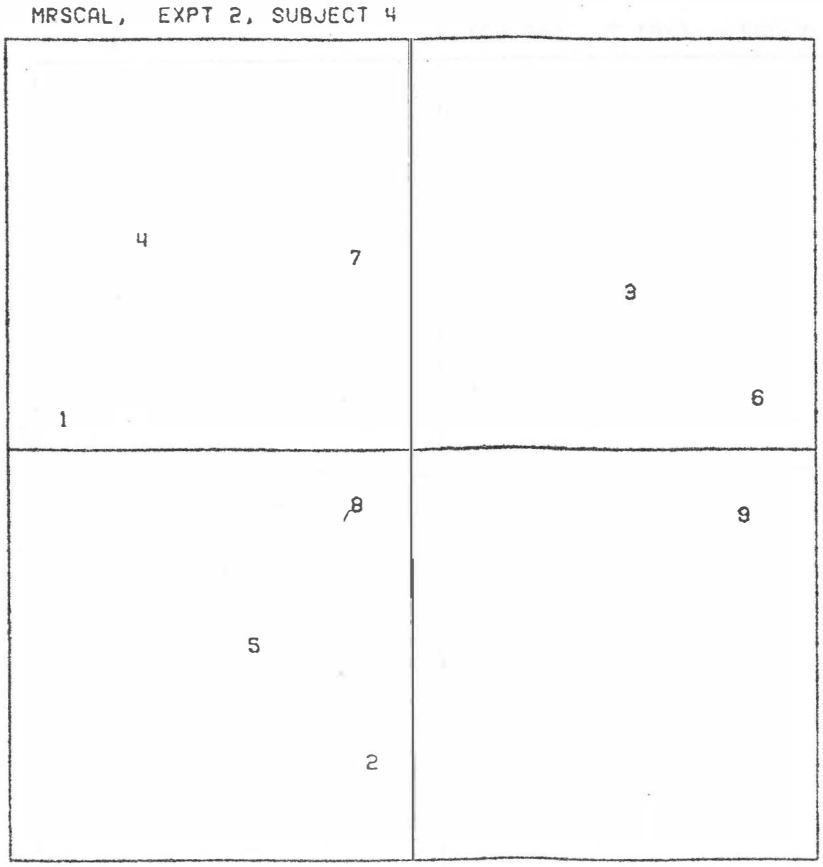

Fig. 9 Similar to Fig. 2.

of an analytic psychophysical listening strategy than of a global speech perception strategy. In a conversational situation one can imagine that a male and a child [œ] are judged identical because the communicative function is the same, illustrated by the use of one phonemic symbol /œ/.

Still a listener is very well able to become aware of the physical differences between these two sounds. This is exactly what the listeners were doing in the present experiment. By just introducing a downward pitch shift the stimuli sounded slightly more natural, but apparently were still judged in an analytic way.

The results are nevertheless surprising in as far as the subjective judgments of the various fundamental frequency and formant steps are concerned: both in an absolute and in a relative sense the difference in fundamental frequency between 140 Hz and 240 Hz is much larger than between 240 Hz and 310 Hz, whereas the subjective dissimilarity shows only a small difference, see for instance Fig. 6. On the other hand, the 26% higher formant values of child versus female formant values are perceptually judged many times more different than the 14% difference in formant values between female and male vowels. A possible explanation for that could be that the male and the female formant values can still be considered to belong to one phoneme class (Pols et al., 1973; van Nierop et al., 1973;

Koopmans-van Beinum, 1980, Fig. 3.11). Moreover, at a psychophysical level, the formant values are less than one critical band apart (Plomp, 1976). In both respects the children's formant values are more extreme, although they reflect average formant values for the actual [œ] of five boys and five girls (Weenink, 1985).
Under the present experimental conditions none of the subjects judged the "standard" male, female, and child [œ] as being identical, they were all very well aware of the physical differences between these signals.

6. ACKNOWLEDGMENTS

The author got much help from Dick van Bergem, Rob Drullman, Irma de Leeuw, and Irenke Meekma in preparing and performing the experiments. Gerard Boxelaar and David Weenink gave software support, while Florien Koopmans-van Beinum was a critical discussant and reader.

7. REFERENCES

Carroll, J.D. and Chang, J.J. (1970), Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition, Psychometrika 35, 283-319.

Koopmans-van Beinum, F.J. (1980), Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions, Ph.D. thesis, University of Amsterdam, p. 54.

Kruskal, J.B. (1964a), Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29, 1-27.

Kruskal, J.B. (1964b), Nonmetric multidimensional scaling: a numerical method, Psychometrika 29, 115-129.

Nierop, D.J.P.J. van, Pols, L.C.W., and Plomp, R. (1973), Frequency analysis of Dutch vowels from 25 female speakers, Acustica 29, 110-118.

Plomp, R. (1976), Aspects of tone sensation. A psychophysical study, Academic Press, London.

Pols, L.C.W. (1977), Spectral analysis and identification of Dutch vowels in monosyllabic words, Ph.D. thesis, Free University, Amsterdam, p. 30.

Pols, L.C.W., Tromp, H.R.C., and Plomp, R. (1973), Frequency analysis of Dutch vowels from 50 male speakers, J. Acoust. Soc. Amer. 53, 1093--1101.

Slawson, A.W. (1968), Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency, J. Acoust. Soc. Amer. 43, 87-101.

Wagenaar, W.A. ans Padmos, P. (1971), Quantitative interpretation of stress in Kruskal's MD-SCALing technique, Br. J. Math. Statist. Psych. 24, 101-110.

Weenink, D.J.M. (1985), this volume, IFA-Proceedings 9.