ORIGINAL ARTICLE

Perceptual scaling of voice identity: common dimensions for different vowels and speakers

Oliver Baumann · Pascal Belin

Abstract The aims of our study were: (1) to determine if the acoustical parameters used by normal subjects to discriminate between different speakers vary when comparisons are made between pairs of two of the same or different vowels, and if they are different for male and female voices; (2) to ask whether individual voices can reasonably be represented as points in a low-dimensional perceptual space such that similarly sounding voices are located close to one another. Subjects were presented with pairs of voices from 16 male and 16 female speakers uttering the three French vowels "a", "i" and "u" and asked to give speaker similarity judgments. Multidimensional analyses of the similarity matrices were performed separately for male and female voices and for three types of comparisons: same vowels, different vowels and overall average. The resulting dimensions were then interpreted a posteriori in terms of relevant acoustical measures. For both male and female voices, a two-dimensional perceptual space was found to be most appropriate, with axes largely corresponding to contributions of the larynx (pitch) and supra-laryngeal vocal tract (formants), mirroring the two largely independent components of source and filter in voice production. These perceptual spaces of male and female voices and their corresponding voice samples are available at: http://vnl.psy.gla.ac.uk section Resources.

O. Baumann · P. Belin
Department of Psychology,
University of Glasgow, Glasgow, UK
e-mail: p.belin@psy.gla.ac.uk

O. Baumann (✉)
Queensland Brain Institute,
The University of Queensland, Brisbane, Australia
e-mail: o.baumann@uq.edu.au

Introduction

The human voice is a very prominent stimulus in our auditory environment as it plays a critical role in most human interactions, particularly as the carrier of speech. Our ability to discriminate and recognize human voices is amongst the most important functions of the human auditory system, especially in the context of speaker identification (Belin, Fecteau & Bédard 2004; van Dommelen, 1990). Theorists have long proposed that speech utterances routinely include acoustic information concerning talker characteristics, in addition to their purely linguistic content. The unique, speaker specific aspects of the voice signal are attributable both to anatomical differences in the vocal structures and to learned differences in the use of the vocal mechanism (Bricker & Pruzansky, 1976; Hecker, 1971), but the nature of the relationship between acoustic output and a listener's perception is yet not fully understood.

One of the first approaches identifying parameters relevant to the perception of interspeaker differences was the application of correlation analysis to the results of evaluative tasks. So-called semantic differential rating scales, which are designed to measure connotative meaning of stimuli (Clarke & Becker, 1969; Holmgren, 1967; Voiers, 1964), as well as rating scales (Clarke & Becker, 1969), have been used to identify speakers or differentiate among voices. Although these studies focused on prosodic features and yielded results to a certain degree inconsistent, it became evident that pitch, intensity and duration are important cues for differentiating voices.

In recent years, several studies have applied multidimensional scaling techniques to listener similarity judgments with the goal to investigate the underlying acoustical parameters. A study by Matsumoto, Hiki, Sone, and Nimura (1973) applied a multidimensional scaling

technique to same-different judgments of pairs of voices uttering five different Japanese vowels and found that the fundamental frequency ($F0$) and formant frequencies accounted for most of the variance in the acoustical measures and were the cues used by the listeners. Walden, Montgomery, Gibeily, Prosek, and Schwartz (1978) conducted a comparable study using similarity judgments of pairs of adult male voices uttering monosyllabic words and derived a four-dimensional perceptual model that correlated with $F0$, word duration, age, and voice qualities rated by speech-language pathologists. Singh and Murry (1978), comparing similarity judgments for adult male and female voices speaking a phrase, found that the gender of the speakers accounted for the major portion of the variance. The second dimension for the male voices was related to $F0$ and the second dimension for female voices was related to duration of the voice sample. They concluded that listeners might attend to different acoustic parameters when judging the similarity of male voices than when judging female voices. The suggestion that the saliency of various acoustic parameters might differ between male and female voices has also been made by other investigators (Aronovitch, 1976; Coleman, 1976). In a follow up study, Murry and Singh (1980) aimed to determine the number and nature of perceptual parameters needed to explain listeners' judgments of similarity for vowels and sentences spoken by male voices compared to female voices. Similarity judgments were submitted to multidimensional analysis via individual differences scaling (INDSCAL) and the resulting dimensions were interpreted in terms of available acoustic measures and one-dimensional voice quality ratings of pitch, breathiness, hoarseness, nasality, and effort. The decisions of the listeners appeared to be influenced by both the sex of the speaker, and whether the stimulus sample was a sustained vowel or a short phrase, although $F0$ was important for all judgments. Aside from the $F0$ dimension, judgments concerning male voices were related to vocal tract parameters, while similarity judgments of female voices were related to perceived glottal as well as vocal tract differences. This finding is corroborated by a study of Hanson (1997), in which the statistical analysis of acoustical parameters of female speech lead to the conclusion that glottal characteristics, in addition to formant frequencies and fundamental frequency, have great importance for describing female speech. Formant structure was apparently important in judging the similarity of vowels for both sexes while perceptual glottal/temporal attributes may have been used as cues in the judgments of phrases (Murray & Singh, 1980).

Kreiman, Gerratt, Precoda and Berke (1992) used separate nonnumeric multidimensional scaling solutions to assess how listeners differ in their judgments of dissimilarity of pairs of voices for the vowel "a". They found in general low correlations between individual listeners, whereby only acoustical parameters that showed substantial variability were perceptually salient across listeners, with naïve listeners mainly relying on $F0$, while expert listeners (speech pathologists and otolaryngologists) also based their judgments on shimmer and formant frequencies.

The aim of our study was to determine if and how the acoustical parameters which are used by normal subjects to discriminate between different speakers, vary if the comparisons are made between a pair of two of the same or two different vowels and whether there is a difference for male and female voices. We further wanted to investigate whether individual voices could be represented as points in a low-dimensional space such that similarly sounding voices would be located close to one another.

By using multidimensional analysis of the average listener similarity judgments and correlating the resulting dimensions with the average acoustic measures over all three vowels for every single speaker we aimed to identify the parameters which were perceptually important across all subjects and voice sets, rather then determining the individual perceptual strategies for every single subject and voice sample. We further conducted a principal component analysis (PCA) on acoustic measures of the used voice samples, to investigate which acoustic parameters form coherent subsets that are relatively independent of one another. This allowed us to compare and discuss the results from this model-free statistical analysis of acoustic measures with the dimensions obtained by multidimensional scaling of perceptual similarity judgments.

Methods

Selection of speakers

Voice samples were recorded from 32 speakers, 16 male and 16 female. For all speakers Canadian French was their native language. The female speakers ranged in age from 19 to 35, with a mean age of 22.5 (SE 1.34) and the male speakers ranged in age from 19 to 40 years, with a mean age of 31.19 (SE 2.61). Each speaker was judged to be free of vocal pathology by one of the experimenters based on informal perceptual judgment, and none of them had received formal voice training. Recordings (16 bit) of the 32 speakers were made in the multi-channel recording studio of Secteur ÉlectroAcoustique in the Faculté de musique, Université de Montréal, using two Bruel & Kjaer 4006 microphones (Bruel & Kjaer; Nærum, DK), a Digidesign 888/24 analog/digital converter and the Pro Tools 6.4 recording software (both Avid Technology; Tewksbury, MA, USA). The lips-to-microphone distance was 120 cm.

Each speaker was instructed to utter the following series of French vowels: "a", "é", "è", "i", "u" and "ou" (in that order) at a comfortable speaking level. The vowels were sustained (about one-second) and produced in isolation (each on a separate breath) in order to minimize list-effects and differences in intonation contours. Recordings of the three vowels "a", "i" and "u" were selected for further acoustical analyses and perceptual similarity judgments.

Procedure

Subjects ($n = 10$, 5 males, 5 females, age range 19–38, mean age 23.9) were presented with all possible pairs of voice samples, with the constraints that a comparison across gender did not occur and that, by random selection, either the AB or BA order of a pair of voices was presented. The order of voice pairs was randomized for each subject as well. In total 4,608 pairs of voice samples were presented to each subject in ten experimental sessions (2,304 pairs for male voices and 2,304 for female voices). The voice samples were presented via a headphone (Beyerdynamic DT 770) and subjects were asked to give a rating regarding how likely they thought it was that the same person spoke both voice samples. To perform their ratings they were presented a visual analogue scale and asked to give their rating by marking an appropriate point onto it. The scale was presented in form of a rectangular box displayed on a computer monitor and they were asked to use a computer mouse to set the marks. The experiment was generated and the response data was collected with the computer programme MCF (Digivox; Montreal, QC, Canada). Subjects were instructed to set a mark on the very left side of the scale, labelled "same" if they were "absolutely sure" that the same person had spoken both voice samples, while they should set a mark on the very right side of the scale, labelled "different" if they were "absolutely sure" that the two voice samples were spoken by two different persons. In the cases were they were not "absolutely sure", they should set a mark on the scale between these two extreme points representing the degree to which they believed that the two voice samples could be spoken by the same person or not. They were told that in all probability it would be rather an exception than the norm that they would be "absolutely sure" about the speaker's identity. They were not told how many different speakers were involved and how many vowel productions each speaker contributed. They were allowed to listen to the voice pairs as often as they wanted before they made their decision. They were also free to make small breaks between trials. The whole experiment consisted of ten sessions of approximately an hour per subject. The sessions were separated by a minimum of six hours and a maximum of four days.

Multi-dimensional scaling (MDS) of similarity judgments

The object of MDS is to reveal relationships among a set of stimuli by representing them in a low-dimensional space so that the distances among the stimuli reflect their relative dissimilarities. To achieve this representation, dissimilarity data arising from a certain number of sources, usually subjects, each relating a certain number of objects pair wise, is modeled by one of a family of MDS procedures to fit distances in some type of space, generally Euclidean or extended Euclidean of low dimensionality.

For both male and female voices, similarity judgments were obtained for 2,304 pairs (16 speakers, 3 vowels) of 2 vowels each. All possible pair combinations were used in the task, including pairs composed of twice the same sound (same vowel by same speaker). The average dissimilarity matrices thus obtained for male voices are displayed in Table 1 and for female voices in Table 2; a value of "0" represents a "same" judgement and a value of "100" a "different" judgement, values in between these two extreme points represent intermediate degrees to which the subjects believed that the two voice samples could be spoken by the same person or not. Multidimensional analyses of the dissimilarity matrices of the two separate groups (female vowel, male vowel) were performed via ALSCAL (SPSS 16.0; SPSS Inc., Chicago, IL, USA), a procedure that has proven useful in the classification of stimuli with obscure perceptual parameters (Carroll & Chang, 1970). The ALSCAL procedure analyzes the perceptual differences between all pairs of speakers as measured by a paired comparison listening task, and provides solutions in a multidimensional space. The resulting dimensions were then interpreted a posteriori by correlating them with acoustical measures that have been reported as relevant for voice recognition (Bachorowski & Owren, 1999; Bruckert, Liénard, Lacroix, Kreutzer, Leboucher, 2006). We refrained from using multiple comparison correction for the correlation analyses, which would be overly conservative, since the several acoustical measures are already known to be not completely independent from each other. For example, Shimmer, Jitter and $F0$ standard deviation have been found to be correlated for sustained vowels (Horii, 1980), and $F0$ and formant frequencies are known to be inherently correlated as well (Singer & Sagayama, 1992).

Acoustic analysis of vowels

Speech sounds are generated by the vocal organs, which are, the lungs, the larynx (containing the vocal cords), the pharynx, the mouth and nasal cavities, and the lungs. The so-called vocal tract is located superior the larynx, and its

Table 1 The average dissimilarity matrix for the 16 male voices (averaged over the three types of vowels), derived from the similarity ratings of 10 subjects

Voice no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	10.35 (4.16)															
2	47.49 (11.11)	15.47 (5.81)														
3	67.37 (12.19)	71.40 (13.14)	11.80 (4.53)													
4	48.58 (10.37)	50.09 (14.22)	70.76 (13.74)	15.98 (6.62)												
5	71.29 (14.21)	63.55 (10.06)	75.78 (12.70)	65.50 (10.22)	13.32 (5.42)											
6	54.94 (12.94)	61.69 (13.57)	75.34 (12.55)	67.02 (12.53)	79.45 (12.96)	13.23 (4.52)										
7	45.08 (11.01)	65.74 (16.79)	59.67 (12.41)	41.19 (16.24)	76.02 (11.71)	58.70 (10.50)	13.00 (5.99)									
8	60.32 (14.45)	74.01 (11.84)	80.90 (11.78)	57.33 (10.65)	76.71 (11.89)	57.31 (10.81)	52.50 (13.23)	7.10 (3.04)								
9	46.81 (11.38)	58.32 (16.35)	65.10 (12.20)	37.43 (14.11)	72.52 (10.06)	55.40 (14.70)	32.33 (13.05)	57.52 (12.60)	10.47 (4.01)							
10	67.39 (12.24)	67.58 (11.65)	50.37 (9.90)	64.83 (7.92)	58.43 (10.60)	57.30 (13.37)	62.71 (10.55)	84.39 (8.74)	64.50 (10.08)	6.45 (2.20)						
11	63.99 (12.54)	70.23 (12.59)	72.20 (11.88)	56.29 (10.22)	74.09 (11.73)	55.82 (13.30)	53.69 (11.66)	67.27 (11.54)	52.42 (10.57)	64.59 (12.24)	26.55 (5.78)					
12	42.59 (10.33)	40.07 (9.58)	64.02 (8.63)	39.31 (11.55)	57.69 (14.82)	51.04 (14.05)	51.47 (11.30)	66.52 (13.48)	41.84 (10.03)	58.32 (11.23)	56.80 (9.85)	11.85 (3.71)				
13	65.85 (14.47)	55.09 (18.25)	58.49 (11.27)	66.29 (11.66)	52.19 (9.12)	74.32 (10.80)	64.99 (14.71)	76.61 (12.05)	70.78 (15.56)	60.26 (8.97)	75.90 (12.69)	59.88 (12.03)	13.63 (6.81)			
14	69.65 (13.92)	65.21 (13.69)	75.11 (14.09)	75.02 (11.49)	56.40 (11.82)	66.59 (15.22)	75.80 (12.67)	81.77 (10.20)	71.03 (11.69)	56.28 (13.17)	72.13 (13.36)	68.41 (11.99)	51.89 (10.31)	15.38 (6.88)		
15	52.16 (11.70)	61.02 (11.57)	73.72 (10.30)	61.76 (13.52)	72.12 (12.35)	42.03 (13.01)	55.68 (10.60)	55.20 (12.01)	53.79 (11.48)	69.93 (10.24)	51.51 (11.26)	61.46 (14.66)	78.23 (10.75)	72.45 (15.10)	23.02 (5.62)	
16	53.62 (10.47)	52.77 (13.20)	56.88 (14.73)	58.15 (9.55)	70.60 (13.42)	59.39 (11.96)	54.98 (11.96)	79.98 (9.87)	60.76 (9.92)	68.30 (11.01)	63.50 (12.01)	51.85 (12.39)	62.67 (13.39)	62.98 (11.28)	67.76 (10.71)	13.99 (7.06)

In brackets the standard deviation is displayed. A value of "0" represent a "same" judgement and a value of 100 a "different" judgement, values in between these two extreme points represent intermediate degrees to which the subjects believed that the two voice samples could be spoken by the same person or not

Table 2 The average dissimilarity matrix for the 16 female voices (averaged over the three types of vowels), derived from the similarity ratings of ten subjects

Voice no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	14.07 (4.38)															
2	53.17 (11.73)	13.43 (6.50)														
3	48.00 (10.32)	65.75 (11.03)	18.73 (7.84)													
4	49.31 (13.60)	55.09 (10.17)	45.92 (12.37)	18.35 (5.23)												
5	57.68 (14.77)	60.24 (13.88)	65.02 (9.30)	54.52 (16.17)	20.58 (5.18)											
6	63.79 (13.52)	64.58 (14.52)	65.50 (9.06)	69.48 (10.82)	53.00 (11.55)	24.08 (9.03)										
7	63.81 (13.02)	63.84 (17.64)	54.93 (11.86)	57.17 (14.55)	76.34 (9.49)	62.21 (13.95)	11.67 (9.36)									
8	45.40 (14.69)	64.08 (14.23)	46.87 (10.49)	52.10 (11.70)	59.77 (13.76)	69.42 (12.21)	60.60 (11.47)	22.25 (8.29)								
9	78.83 (7.92)	72.35 (16.13)	68.63 (8.58)	71.12 (10.75)	66.06 (6.67)	55.63 (15.79)	78.71 (9.28)	63.66 (17.78)	8.55 (5.94)							
10	65.11 (10.88)	64.56 (10.78)	57.93 (10.24)	72.61 (8.73)	48.75 (9.12)	61.51 (10.55)	71.35 (10.92)	63.07 (10.10)	53.93 (9.75)	12.42 (7.99)						
11	76.97 (13.22)	85.32 (7.93)	69.50 (12.47)	76.77 (12.75)	75.39 (10.54)	69.81 (9.29)	75.78 (9.61)	62.85 (14.69)	31.81 (10.46)	66.89 (8.73)	16.03 (6.13)					
12	57.10 (9.56)	67.31 (10.99)	39.72 (6.33)	49.02 (10.04)	67.30 (11.57)	63.31 (11.82)	52.95 (10.02)	57.54 (14.47)	67.60 (9.96)	66.61 (8.09)	63.05 (15.54)	8.47 (3.99)				
13	54.76 (9.96)	58.59 (13.29)	49.68 (12.11)	53.96 (10.45)	66.22 (10.77)	73.28 (11.34)	49.31 (11.30)	68.69 (12.38)	76.58 (11.38)	67.01 (10.97)	74.41 (11.63)	57.89 (10.19)	9.87 (5.27)			
14	48.98 (9.13)	68.36 (9.93)	66.46 (9.08)	49.79 (7.91)	68.01 (12.85)	67.37 (9.17)	58.21 (16.47)	54.20 (12.81)	58.68 (11.20)	59.97 (13.68)	64.27 (9.28)	49.36 (9.58)	76.28 (11.07)	14.17 (5.84)		
15	65.78 (12.45)	71.98 (11.81)	61.03 (8.64)	63.19 (6.42)	70.02 (8.74)	59.99 (12.32)	70.74 (10.88)	49.41 (12.59)	37.63 (12.59)	57.60 (10.47)	49.58 (11.03)	38.30 (9.25)	74.76 (10.67)	54.86 (10.08)	21.38 (5.36)	
16	49.99 (10.75)	54.22 (8.50)	45.99 (5.67)	42.51 (12.96)	53.94 (14.09)	54.70 (12.57)	68.77 (11.26)	57.78 (13.57)	49.19 (14.87)	44.85 (6.28)	67.33 (9.72)	55.74 (8.10)	66.63 (9.18)	46.47 (14.47)	55.95 (7.57)	13.00 (6.82)

In brackets the standard deviation is displayed. A value of "0" represent a "same" judgement and a value of 100 a "different" judgement, values in between these two extreme points represent intermediate degrees to which the subjects believed that the two voice samples could be spoken by the same person or not

shape is varied extensively by movements of the tongue, the lips and the jaw. The space between the vocal folds is called the glottis; the vocal folds can open and close, varying thereby its size, which in turn affects the flow of air from the lungs. The source-filter theory describes speech production as a process of two largely independent stages, involving the generation of a sound source, with its own spectral shape and spectral fine structure, which is then shaped or filtered by the resonant properties of the vocal tract. The term "glottal source" refers to the sound energy produced by the flow of air from the lungs past the vocal folds as they open and close quite rapidly in a periodic or quasi-periodic manner. The sound energy produced by the vocal folds by modulating the airflow from the lungs is a periodic complex tone with a relatively low fundamental frequency, also referred to as the fundamental frequency of phonation ($F0$). The vocal tract subsequently filters the produced sound, introducing resonances (called formants) at certain frequencies. The formants are numbered; with the one with the lowest frequency called the first formant ($F1$), the next the second formant ($F2$), and so on. The centre frequencies of the formants differ with the shape of the vocal tract. Vowels are the speech sounds that are characterized most easily, since their formants and other acoustic features are relatively stable over time, when spoken in isolation (Moore, 2003).

Because we wanted to get a general measure of vocal range, we used means for vocal measurements across the three vowels, which is more representative of a speaker's vocalizations and reduces statistical dispersion. We used PRAAT 4.2.07 software (P. Boersma and D. Weenink, http://www.praat.org) to measure mean $F0$, between the three vowels; the overall temporal variation of $F0$ ("$F0$-SD" in the tables), as the standard deviation of $F0$ over the entire voice sample, which gave us an indicator for the intonation; the jitter, a measure of local frequency variation of the $F0$, as the average absolute difference between consecutive periods, divided by the average period; as well as shimmer, a measure of local amplitude variation, as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude. We measured the peak frequencies averaged across the whole stimulus duration of the first five formants ($F1$–$F5$) of each vowel and then calculated their means across the three vowels (FFT spectrum, Fourier method, all parameters were default values recommended by the authors of PRAAT; except the maximum formant frequency for female voices, which was set to 6,500): 5-ms Gaussian window, 2-ms time step, 20 Hz frequency step, 50 dB dynamic range, 5,000 Hz maximum formant frequency. Overall formant dispersion was calculated ("Disp $F1$–$F5$" in the tables), as the mean interval between formant frequencies, for each vowel, and the overall formant

dispersion across the three vowels. Further, the overall formant dispersion was calculated with only the fourth and fifth formant ("Disp $F4$–$F5$" in the tables) because these two formants are less likely to be dependent on the kind of vowel (Fant, 1960); this parameter was measured in previous studies (Collins 2000; Collins & Missing, 2003). Using Praat we further calculated the harmonics to noise ratio in dB ("HTN" in the tables) of each voice sample, the degree of acoustic periodicity, which reflects the hoarseness of a sound (Yumoto, Sasaki & Okamura, 1984), and the duration ("Dur" in the tables) of the voice samples. Finally we conducted a loudness matching experiment, where the subjects had to adjust the intensity of every voice sample (in steps of ± 1 dB) until it seemed equal in loudness to a standard voice sample, which was not used in the experiment. We then used the relative differences in dB relative to the standard voice sample as the measure of loudness.

Principal component analysis

Principal component analysis (PCA) is a statistical technique applied to a set of variables with the aim to reduce the original set of variables and to reveal which variables in the set form coherent subsets that are relatively independent of one another. Variables that are correlated with one another but largely independent of other subsets of variables are combined into components. Thereby the components are thought to reflect underlying processes that have created the correlations among variables (Tabachnick & Fidell, 1996). The results of a PCA are usually discussed in terms of the variance explained by each component and the component loadings. The loadings can be understood as the weights for each original variable when calculating the principal component, or as the correlation of each component with each variable.

We conducted a PCA with the vocal parameters of the voice samples from each subject (averaged over all three types of vowels), to reduce the large set of acoustical parameters to a small number of components, and to compare these to the results obtained from the MDS. This allowed us to investigate the importance of specific acoustic parameters for differentiating speakers, in human observers as compared to the outcome of a model-free statistical technique.

Results

Principal components of acoustical measures

Principal components analyses (PCA) (SPSS 16.0; SPSS Inc., Chicago, IL, USA) with varimax rotation were

conducted in order to examine clustering among variables. These PCA were conducted separately for males and females because of the large differences in $F0$ and formant frequencies. The analysis was restricted to a 2 factorial solution to be directly comparable to the 2 dimensional constellation of the perceptual space derived from the MDS procedure. The resulting solutions accounted for 49.09 and 46.34% of the cumulative variance for males and females, respectively. For males the first factor (28.43%) corresponded to jitter, shimmer and the standard deviation of $F0$, and inversely to duration, while the second factor (20.66%) corresponded best to the $F5$, the dispersion between $F1$ and $F5$, and the dispersion between $F4$ and $F5$ (see Table 3). For females the first factor (24.97%) was correlated to $F5$, the dispersion between $F1$ and $F5$, and the dispersion between $F4$ and $F5$. The second factor (21.37% correlated highly with shimmer and jitter, and inversely with duration (see Table 4).

Multidimensional analysis and construction of the voice space

Multidimensional analyses of the similarity matrices were performed separately for male and female voices and for three types of comparisons: same vowels, different vowels and overall average. For each of the two groups and all types of comparisons studied, a two-dimensional solution was found to be most appropriate, based on the criteria of interpretability, uniqueness, and percentage of accounted-

Table 3 Results of the PCA for the male voices (averaged over the three types of vowels)

Component		
	1	2
$F1$	0.517	−0.245
$F2$	−0.382	0.242
$F3$	0.484	−0.231
$F4$	0.162	0.052
$F5$	0.246	*0.910*
$F0$	0.543	0.375
$F0$-SD	*0.797*	0.165
Dur	*−0.825*	−0.014
Disp ($F1$–$F5$)	0.185	*0.911*
Disp ($F4$–$F5$)	0.108	*0.817*
Shimmer	*0.798*	−0.319
Jitter	*0.910*	−0.254
Loudness	0.084	0.435
HTN	0.208	−0.275

Rotated component loadings for principal components extraction with varimax rotation. A cutoff point of ±0.75 was used to include a variable in a component, and variables meeting this criterion are noted in italics

Table 4 Results of the PCA for the female voices (averaged over the three types of vowels)

Component		
	1	2
$F1$	0.130	−0.199
$F2$	0.241	0.468
$F3$	−0.493	0.300
$F4$	0.533	−0.034
$F5$	*0.973*	0.130
$F0$	−0.236	−0.089
$F0$-SD	0.029	0.568
Dur	−0.021	*−0.714*
Disp ($F1$–$F5$)	*0.969*	0.148
Disp ($F4$–$F5$)	*0.823*	0.178
Shimmer	0.216	*0.835*
Jitter	−0.109	*0.840*
Loudness	−0.433	0.399
HTN	−0.131	0.348

Rotated component loadings for principal components extraction with varimax rotation. A cutoff point of ±0.75 was used to include a variable in a component, and variables meeting this criterion are noted in italics

for variance. The ALSCAL results were interpreted by plotting and examining the dimensions and by examining correlations between each of the dimensions and the available acoustic measures. The significant correlation coefficients ($P < 0.05$) between the two ALSCAL dimensions and the acoustic measures for each of the two groups are presented in the Tables 5, 6, 7, 8. The 2-dimensional ALSCAL solutions for each of the groups are graphically represented in Figs. 1 and 2. Suggested interpretations for each dimension are indicated on the figures.

For the male voices (averaged over all types of comparisons) the overall model fit for a two-dimensional solution had a Stress value of 0.19403 and a squared correlation value (RSQ) of 0.81215. According to Borg & Staufenbiel (1989) Stress values <0.2 constitute a sufficient fit, therefore we did not calculate a three dimensional model. The first axis of this model correlated only with the $F0$ (Sig. (2-tailed) 0.000 Pearson Correlation −0.832). For the two models taking only same or different vowels into account the first axis correlated strongest with the $F0$ as well (different vowels Sig. (2-tailed) 0.000; Pearson Correlation −0.797; same vowels Sig. (2-tailed) 0.000; Pearson Correlation −0.817). The second axis correlated highest with the formant dispersion between $F4$ and $F5$ (Sig. (2-tailed) 0.004; Pearson correlation −0.680), and $F4$ (Sig. (2-tailed) 0.007; Pearson correlation 0.649) (see Table 5). A similar pattern was evident for the models taking only pairs of different vowels or same vowels into account (see Table 6). The model fit for the

Table 5 Pearson correlation coefficients between the 2 axes of the perceptual space and the acoustical parameters for the male voices (averaged over all types of comparisons and vowels)

	Dim1	Dim2	$F1$	$F5$	$F0$-SD	Dur	Disp ($F1$–$F5$)	Shimmer	Jitter
$F1$	−0.577(*)								
$F4$		0.649(**)							
$F0$	−0.832(**)		0.524(*)						
Dur					−0.599(*)				
Disp ($F1$–$F5$)				0.995(**)					
Disp ($F4$–$F5$)		−0.680(**)		0.690(**)			0.692(**)		
Shimmer					0.665(**)	−0.689(**)			
Jitter					0.761(**)	−0.743(**)		0.916(**)	
Mean ($F1$–$F4$)									−0.529(*)

Only significant correlations are displayed

* Correlation is significant at the 0.05 level (2-tailed)

** Correlation is significant at the 0.01 level (2-tailed)

Table 6 Pearson correlation coefficients between the 2 axes of the perceptual space and the acoustical parameters for the male voices (only significant correlations are displayed)

	Dim1	Dim2
Taking only comparisons between different vowels into account		
$F1$	−0.499(*)	
$F3$		0.512(*)
$F4$		0.645(**)
$F0$	−0.797(**)	
Disp ($F4$–$F5$)		−0.653(**)
Shimmer		0.528(*)
Taking only comparisons between same vowels into account		
$F1$	−0.570(*)	
$F4$		0.604(*)
$F0$	−0.817(**)	
$F0$-SD		−0.530(*)
Disp ($F4$–$F5$)		−0.696(**)
Loudness	−0.520(*)	

* Correlation is significant at the 0.05 level (2-tailed)

** Correlation is significant at the 0.01 level (2-tailed)

model with only different vowels was not as good (Stress = 0.20309; RSQ = 0.77468) as the average model for all types of comparisons. The same was true for the model taking only same vowels into account (Stress = 0.21682; RSQ = 0.72560). This shows that collapsing the similarity rating over same and different vowel judgements is a viable approach, which increases the model fit.

For the female voices (averaged over all types of comparisons), the overall model fit for a two dimensional solution had a Stress value of 0.18612 and a RSQ of 0.78466. The first axis of this model correlated only with the $F0$ (Sig. (2-tailed) 0.000; Pearson correlation −0.875).

For the two models taking only same or different vowels into account the axis correlated strongest in both instances with the $F0$ as well (different vowels: Sig. (2-tailed) 0.000; Pearson correlation −0.906; same vowels: Sig. (2-tailed) 0.000; Pearson correlation −0.809). The second axis in the model averaged over all types of comparisons correlated highest with $F1$ (Sig. (2-tailed) 0.007; Pearson correlation 0.642). In the models taking only different vowels into account the axis correlated strongest with the $F1$ (Sig. (2-tailed) 0.002; Pearson correlation 0.709) and jitter (Sig. (2-tailed) 0.012; Pearson correlation −0.613), and for the model only taking same vowels into account the second axis correlated best with jitter (Sig. (2-tailed) 0.007; Pearson correlation −0.645) and the $F1$ (Sig. (2-tailed) 0.36; Pearson correlation 0.527). (see Tables 7, 8 for details). The model fit for the model with only different vowels was not as good (Stress = 0.19746; RSQ = 0.79907) as that for the average of all types of comparisons. The same was true for the model taking only same vowels into account (Stress = 0.25753; RSQ = 0.6559). As for the male voices, collapsing the similarity ratings over same and different vowel judgements increased the model fit. It is worth mentioning that the subjects were not using the duration of the voice samples for their similarity ratings, even though the duration of the voice samples had very high components loadings in the PCA for both the female and male voices (see Tables 3, 4).

Discussion

The purpose of our study was to determine which acoustical parameters normal subjects use to discriminate between different speakers, whether these parameters vary

Table 7 Pearson correlation coefficients between the 2 axes of the perceptual space and the acoustical parameters for the female voices (averaged over all types of comparisons and vowels)

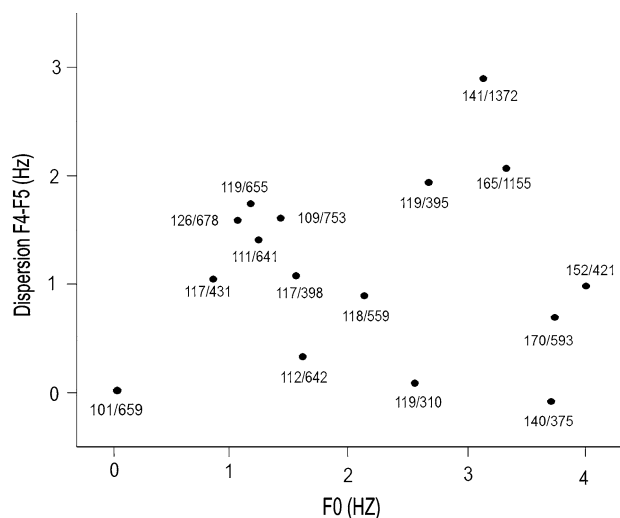|	Dim1	Dim2	F3	F4	F5	F0	Dur	Disp (F1–F5)	Shimmer	Jitter
F1		0.642(**)								
F5				0.554(*)						
F0	−0.875(**)									
F0-SD		−0.561(*)				0.522(*)				
Disp (F1–F5)				0.529(*)	0.996(**)					
Disp (F4–F5)			−0.534(*)		0.843(**)			0.855(**)		
Shimmer							−0.589(*)			
Jitter		−0.580(*)					−0.499(*)		0.655(**)	
Mean (F1–F4)										−0.542(*)

Only significant correlations are displayed

* Correlation is significant at the 0.05 level (2-tailed)

** Correlation is significant at the 0.01 level (2-tailed)

Table 8 Pearson correlation coefficients between the 2 axes of the perceptual space and the acoustical parameters for the female voices (only significant correlations are displayed)

	Dim1	Dim2
Taking only comparisons between different vowels into account		
F1		0.709(**)
F0	−0.906(**)	
F0-SD	−0.499(*)	
Jitter		−0.613(*)
Taking only comparisons between same vowels into account		
F1		0.527(*)
F0	−0.809(**)	
F0-SD		−0.561(*)
Jitter		−0.645(**)

* Correlation is significant at the 0.05 level (2-tailed)

** Correlation is significant at the 0.01 level (2-tailed)

Fig. 1 The two-dimensional voice space: a spatial model derived with the ALSCAL procedure from dissimilarity ratings on 16 male voices by 10 subjects (averaged over all types of comparisons and vowels). The acoustic correlates of the perceptual dimensions are indicated with arbitrary units. For each voice sample the average F0 and formant dispersion between F4 and F5 are indicated

when the comparisons are made between pairs of two of the same or different vowels, and if there is a difference for male and female voices. We further wanted to investigate if individual voices could be represented as points in a low-dimensional space such that similarly sounding voices would be located close to one another.

In total 4,608 pairs of voice samples were presented to each subject in ten experimental sessions, which is around four to seven times the amount of comparisons used in previous studies to measure the similarity between speakers (Kreiman et al., 1992; Matsumoto et al., 1973).

Previous reports have suggested that the acoustic attributes used to distinguish among individual speakers are different for male and females—aside from the F0 dimensions judgments concerning male voices were related to vocal tract parameters, while similarity judgments of

female voices were related to perceived glottal and vocal tract difference—(Murry & Singh, 1980; Singh & Murry, 1978). In contrast our data suggested a more similar pattern across sexes, while our finding that F0 is a primary parameter for differentiating among speakers is consistent with previous studies (Clarke & Becker, 1969; Holmgren, 1967; Murry & Singh, 1980; Singh & Murry, 1978; Voiers, 1964; Walden et al., 1978). For male and female voices, F0 appears to be the primary dimension for judgments of sustained vowels. This is in concordance with Kreiman et al. (1992), who found that naïve listeners perceived normal voices (producing the vowel "a") in terms of F0.
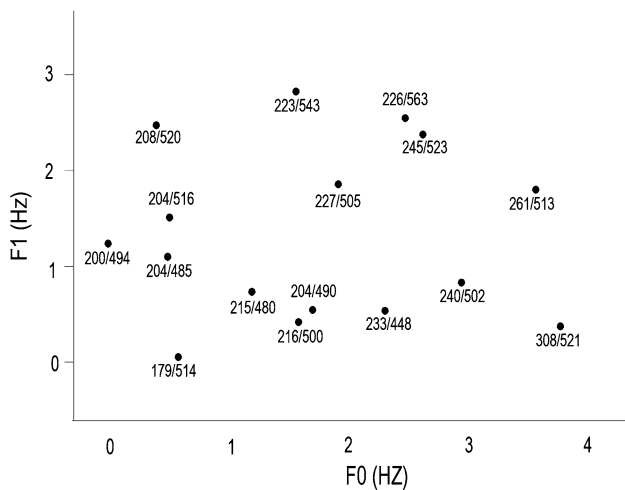
Fig. 2 The two-dimensional voice space: a spatial model derived with the ALSCAL procedure from dissimilarity ratings on 16 female voices by 10 subjects (averaged over all types of comparisons and vowels). The acoustic correlates of the perceptual dimensions are indicated with arbitrary units. For each voice sample the average $F0$ and $F1$ are indicated

Regarding the second dimension, for differentiating female voices the $F1$ was of greater importance while it was for males the dispersion between $F4$ and $F5$ (and to a similar degree the $F4$ alone as well). The $F4$ and $F5$ are known to be more independent from the spoken vowel (Fant, 1960), but they have as well typically much less energy in female voice spectrograms compared to male voices. So even though the $F4$ and $F5$ would be more suitable for classifying talkers, the energy level could in most cases be just too low to be used to identify female speakers.

Overall, the two axes of the obtained perceptual space of voices largely represented contributions of the larynx and supra-laryngeal vocal tract, which, according to the source-filter theory, are largely independent components of voice production. According to the results from the PCA the $F0$, relative to other measures, did not have a very high loading on the two principal factors, which leads to the conclusion that humans might rely to a large extent on an acoustical parameter, which from a signal processing point of view is not very informative to differentiate between speakers. According to the PCA results it would be a better strategy to use shimmer, jitter, the standard deviation of $F0$, $F5$, the dispersion between $F1$ and $F5$, the dispersion between $F4$ and $F5$, or the duration of the voice samples, to differentiate among the talkers. This assumption is supported by studies like Bachorowski and Owren (1999) who were using statistical discriminant classifications of individual talker identity and found that the formant frequency variables correctly classified 42.7% of cases. In contrast, the $F0$ resulted in correct classification of only 13.3% in males and

7.4% of cases in females. Given the fact that the observers in the present experiment on average classified 70.18% of the voice samples correctly, the ability of (naive) human observers appears to be far from perfect in classifying speaker identity, using single vowels uttered by unfamiliar speakers. But it should be mentioned that even pure statistical classifications of single vowels are not able to achieve perfect results, e.g. in the study of Bachorowski and Owren (1999) only in 75.6% voice samples the speaker identity was correctly identified. In real-life situations humans may also rely more on features like intonation of the sentence, typical phrases, construction of sentences, richness of the voice and dialects; variables which are difficult to measure and occur over time scales (Endres, Bambach & Flosser, 1971) larger than the duration of a vowel.

Another reason for the relatively low-level of performance might be the fact that non-familiar speakers spoke the voice samples. If a subject would be trained with several voice samples of the same speaker it would allow the formation of a more versatile representation of its characteristics, which could lead to a much better accuracy in a voice discrimination task.

The level of experience is also an important factor. In the study of Kreiman et al. (1992) where expert and naïve listeners were asked to give similarity ratings for speakers uttering the vowel "a" it became evident that while naïve listeners relied mostly on $F0$, experts relied as well on formants and shimmer to make their judgments.

Overall, the perceptual space obtained from MDS of similarity ratings appears to roughly correspond to a separation of the contributions of the source and filter parts of the vocal apparatus. This is a plausible interpretation, since the source-filter theory proposes that these two components of voice production are largely independent. Thus, despite the overemphasis on $F0$, it seems that the perceptual system makes a good use of the information provided in the voice samples.

In conclusion, we found that a simple two-dimensional space seems to be an appropriate and sufficient representation of perceived speaker similarity. The 'voice space' derived by us can be a useful as foundation for future experiments on voice perception and therefore a valuable contribution to the community of voice researchers. The obtained perceptual spaces of male and female voices and their corresponding voice samples are available at: http://vnl.psy.gla.ac.uk section Resources.

References

Aronovitch, D. S. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of Social Psychology, 99*, 207–220.

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America, 106*, 1054–1063.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Science, 8*, 129–135.

Borg, I., & Staufenbiel, T. (1989). *Theorien und Methoden der Skalierung*. Bern: Huber.

Bricker, P. D., & Pruzansky, S. (1976). Speaker recognition. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 295–326). New York: Academic.

Bruckert, L., Liénard, J. S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. In: *Proceedings of the royal society. Biological sciences* (Vol. 273, pp. 83–89).

Carroll, J. D., & Chang, J. (1970). An analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrica, 35*, 283–319.

Clarke, F. R., & Becker, R. W. (1969). Comparison of techniques for discriminating among talkers. *Journal of Speech and Hearing Research, 12*, 747–762.

Coleman, R. O. (1976). A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech and Hearing Research, 19*, 168–180.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour, 40*, 773–780.

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour, 65*, 997–1004.

Endres, W., Bambach, W., & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *The Journal of the Acoustical Society of America, 49*, 1842–1848.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton & Co.

Hanson, H. (1997). Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America, 101*, 466–481.

Hecker, M. H. L. (1971). Speaker recognition: An interpretive survey of the literature. ASHA Monographs No. 16

Holmgren, G. (1967). Physical and psychological correlates of speaker recognition. *Journal of Speech and Hearing Reserch, 10*, 57–66.

Horii, Y. (1980). Vocal shimmer in sustained phonation. *Journal of Speech and Hearing Research, 23*, 202–209.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35*, 512–520.

Matsumoto, H., Hiki, S., Sone, T., & Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics, 21*, 428–436.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing*. Amsterdam: Academic Press.

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America, 68*, 1294–1300.

Singer, H., & Sagayama, S. (1992). Pitch dependent phone modelling for HMM based speech recognition. *Acoustics, Speech, and Signal Processing, 1*, 273–276.

Singh, S., & Murry, T. (1978). Multidimensional classification of normal voice qualities. *The Journal of the Acoustical Society of America, 64*, 81–87.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. New York: HarperCollins.

van Dommelen, W. A. (1990). Acoustic parameters in human speaker recognition. *Language and Speech, 33*, 259–272.

Voiers, W. D. (1964). Perceptual bases of speaker identity. *The Journal of the Acoustical Society of America, 36*, 1065–1073.

Walden, B. E., Montgomery, A. A., Gibeily, G. T., Prosek, R. A., & Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech and Hearing Research, 21*, 265–275.

Yumoto, E., Sasaki, Y., & Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *Journal of Speech and Hearing Research, 27*, 2–6.