TR－A－0132 14

Perceptual aspects of voice individuality

上田 和夫

1992. 1.28

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町光台2-2 ☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

Perceptual aspects of voice individuality

Kazuo Ueda

Hearing & Speech Perception Department
ATR Auditory and Visual Perception Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan

Studies on some perceptual aspects of voice individuality, which include sex identification, voice quality, and speaker identification perception, were reviewed and discussed. The main conclusion was that each voice is to be viewed as a unique pattern which contains many different acoustic parameters, and that listeners utilize different parameters for different situation.

INTRODUCTION

This article reviews some perceptual aspects of voice individuality. Following the pioneering normalization study by Peterson and Barney (1952), there were numerous studies of voice individuality marked by various approaches. Normalization should take place whenever some "irrelevant" variations exist. It is some form of these variations that, in turn, characterizes voice individuality (Peterson and Barney, 1952).

In the following, we will discuss the general methodology and terminology of voice individuality research, research on sex identification, voice quality, and speaker identification with familiar voices. The research on sex identification will be introduced more precisely than the other research.

I. METHODOLOGY AND TERMINOLOGY

It is a matter of course that the reliability of evidence gained from a given experiment is largely dependent upon what specific methodology has been employed, if one is aiming to say something in terms of science: the right method should be employed for the right purpose, to the right stimuli with the right presentation, and so on. The interpretation of the results is also dependent on the method.

Some of the methodology which has been employed in voice individuality research will be reviewed in this section. Some of these methodology will be criticized. In addition, the terminology problem concerning the distinction between discrimination and identification will be discussed.

A. Problems of the methodology utilized in the studies on sex identification

Among the methodologies which will be introduced in section, II., there are two about which I would like to make negative comments in advance.

First, the methodology utilizing synthetic speech stimulus poses the problem that perfect separation of the glottal source characteristics and the vocal tract resonance characteristics is generally impossible. It also poses the problem of naturalness.

Second, the methodology utilizing the electrolarynx has many problems of uncertainty (Kobayashi, 1991): the sound differs according to the condition of skin attachment; it also differs according to the thickness of the skin; since the sound source is, in effect, located somewhere in the middle of the vocal tract, the effective vocal tract length is ambiguous; a glottal opening results in an alternation of the vocal tract resonance characteristics, but there is no way to check the degree of glottal closure (Coleman, 1976).

Thus, if it is at all possible, I would prefer to use natural speech rather than synthetic or speech produced using an electrolarynx.

B. Approaches to underlying perceptual structure: Factor analysis and multidimensional scaling analysis

These approaches are most appropriate to find the fundamental structure of a psychological space. If the nature of the research requires any fine statistical discrimination among conditions, these approaches are inappropriate.

For factor analysis, raw data is obtained from semantic differential (SD) ratings, with discrete presentation of each stimulus in most cases, whereas for multidimensional scaling (MDS) analysis, the data is often obtained from similarity (or dissimilarity) ratings for paired stimuli, or from forced choice of the most similar pair in three stimuli presented successively. Since the psychological distance obtained in this way should be regarded as rank-order level, non-metric MDS techniques, such as Kruskal's, INDSCAL, or ALSCAL, have been preferably applied.

To obtain reliable results, factor analysis is thought to need a fairly large number of subjects (say, 100), because individual differences among subjects in SD ratings are generally large, and the differences are subject to systematic bias depending on the subject groups (Iwashita, 1979). Further, the choice of SD scales, i.e., the choice of adjectives, is a crucial factor in making the measurement meaningful. If the experimenter drops any perceptually important adjectives, and hence SD scales, it may lead to nonsense results.

MDS, on the other hand, which is usually used for analyzing dissimilarity, i.e., psychological distance, between paired stimuli, is less apt to drop fundamental dimensions, if any. Dissimilarity data are regarded to contain all the perceptual aspects that enable us to distinguish each stimulus (cf. Schiffman et al., 1981).

The choice of stimuli is one of the crucial points in both methods. Appropriate stimulus variation is necessary to guarantee the generality of the results (perceptual structure). It should also be noted that the frame of reference is influenced by the stimulus range or context (cf., for example, Mullennix et al., 1991).

Essentially, two stimuli are considered to have been distinguished if the SD rating profiles of the stimuli are different, or the dissimilarity between them is not zero. Otherwise, the stimuli are supposed to be identically perceived. In this sense, these measures may be regarded as certain variations of the discrimination task.

Some of the studies utilizing MDS analysis have employed multiple correlation (or regression) analysis to find correspondences between perceptual and physical dimensions.

C. Discrimination and identification

There are several researchers who make no distinction between discrimination and identification (cf., Shearme and Holmes, 1959; Coleman, 1973; Itoh and Saito, 1986). However, it is of vital importance to make the distinction. There are three reasons which support the distinction: Two are theoretical considerations, and the other is based on neuropsychological evidence.

From the theoretical point of view, first of all, the discrimination task can be performed no matter what perceptible difference exists between the two stimuli, whereas the identification task cannot necessarily be performed even if there is a perceptible difference.

In the second reason, which is somewhat related to the first, the distinction should be made because the extent to which long-term memory is involved should be significantly different according to the task required. Since the identification task involves name verification in the long-term memory and name retrieval from long-term memory, whereas the discrimination task only requires finding a difference between the two kinds of information stored in short-term memory, the identification task imposes a much heavier load on long-term memory than does the discrimination task (see also, Abberton and Fourcin, 1978; Van Lancker, Kreiman, and Emmorey, 1985).

The last point is based on a neuropsychological fact which shows that recognizing a familiar voice and discriminating among unfamiliar voices may be selectively impaired in brain-damaged subjects (Van Lancker and Kreiman, 1987). This implies that identification involves some specific functions of a certain part of the brain which is not used in discrimination. Thus, these tasks are to be distinguished from this point of view also.

Taken together, these two tasks are essentially different in their nature, and must not be confused. Thus, the distinction will be strictly adhered to in this article.

II. SEX IDENTIFICATION

It has been shown that the most psychologically prominent dimension of voice quality is sex difference (Singh and Murry, 1978; see III). It is possible to consider that the difference is partly caused by the differences observed in formant frequencies of vowels; Females as a group have higher average formant frequencies than males (Peterson and Barney, 1952). Speaker sex identification has been studied for voiceless fricatives, whispered vowels, voiced vowels, and vowels produced by an electrolarynx. Though there have been several studies which focus on sex identification in children's voices (e. g., Bennett and Weinberg, 1979), we will restrict ourselves to adult voices here.

Voiceless fricatives were employed in the studies of Schwartz (1968) and Ingemann (1968). Eighteen speakers (nine females and nine males) and 10 listeners participated in Schwartz's study, and 14 to 15 phonetically trained speakers (six to seven females and eight males) and 10 listeners participated in Ingemann's study. The duration of the stimuli was 500 ms (the stimuli were excised from the central portion of each phonation) in Schwartz's study, whereas it varied from less than 1 s to over 3.5 s in Ingemann's study. They found that speaker sex could be accurately identified, especially /h/, /ʃ/, and /s/: more than 90% correct identifications were achieved for these three fricatives. From spectrographic analysis for /ʃ/ and /s/, Schwartz found that the female spectra generally tended to be higher in frequency than that of the male. Moreover, Ingemann concluded that as the portion of the vocal tract in front of the constriction diminishes, so too does identification of the speaker's sex. These results, together with the fact that there was no laryngeal fundamental available for these voiceless consonants, suggest that accurate speaker sex identification is possible from vocal tract resonance information alone.

Furthermore, the same conclusion was reached from an experiment that dealt with excised (400 ms), whispered vowels (Schwartz and Rine, 1968). Schwartz and Rine studied sex

identification for whispered vowels /i/ and /a/. Ten speakers (five male and five female) and eight listeners were employed. They found that 100% correct identification was achieved for /a/, and that 95% correct identification was achieved for /i/. Similar spectrum shifts as in the fricatives were observed for female utterances.

To eliminate between-subject variations in laryngeal fundamental frequency, Coleman (1971) utilized a single-frequency (85 Hz) electrolarynx in his study on male and female voice quality and its relationship to vowel formant frequencies. Speech samples (a prose passage, and vowels /i/ and /u/) obtained from 20 speakers (10 females and 10 males) were presented to 15 listeners who were asked to identify the sex of each speaker. Correct identifications were made 88% of the time. From spectrographic analysis, the means of the first three formant frequencies in each vowel were obtained. The mean frequencies were closely correlated with the degree (subjective certainty) of male or female quality of the voices. Since an electrolarynx was used as a sound source, the characteristics of the individual vocal tract resonances were considered to involve a cue to judge sex identification of the speakers.

In this study, it was observed that it was easier for the judges to identify the male speakers. Coleman suggested that one explanation for this may come from the frequency of the sound produced by the electrolarynx. According to him, "while 85 Hz is well below the average vocal fundamental of either sex it is closer to that of the males. ...the low frequency of the electrolarynx may have given a male quality to the voice..."

Later, Coleman (1976) compared the contributions of the fundamental frequency (F_0) and vocal tract resonances to perception of maleness and femaleness. In his first experiment, a total of 40 speakers, 20 females and 20 males, uttered several passages and four vowels normally. To eliminate the influence of suprasegmental differences between the sexes as much as possible, the speech segments were played backward to a group of 17 listeners. The listeners were asked to determine the sex of each speaker, and to estimate the degree of femaleness and maleness of each voice on a seven-point scale. Rank-order correlation

coefficients were computed between average frequencies of the first three formants, fundamental frequencies, and ratings of femaleness and maleness. The highest correlation coefficient of 0.94 was attained between $F0$ and ratings of femaleness or maleness. Thus, he concluded that "...listeners were basing their judgments of the degree of maleness or femaleness in the voice on the frequency of the laryngeal fundamental."

In the second experiment, he used a laryngeal vibrator as a substitute for the sound source for a normal glottal tone. Two frequencies of the vibrator were chosen to represent both males (120 Hz) and females (240 Hz). Thus, there were two conditions when the stimuli were produced: In one condition, $F0$ and the vocal characteristics of the speakers were consistent, while in the other condition, they were inconsistent. Ten speakers (five females and five males) were selected on the basis of the vocal tract resonance measure in the first experiment: the females with the highest vocal tract resonances and the males with the lowest vocal tract resonances were selected. The 25 listeners were asked to identify the sex of the speakers, and the results showed that the identification was strongly biased towards the male side, indicating that a "female" $F0$ was an ineffective cue to the femaleness of the voice when it was combined with the male vocal tract resonance features.

Another experiment, which examined the relative importance of the laryngeal fundamental frequency and vocal tract resonance characteristics in speaker sex identification tasks, was conducted by Lass et al. (1976). A total of 20 speakers (10 females and 10 males) produced six sustained isolated vowels (/i, ɛ, æ, ɑ, o, u/). All vowels were recorded under normal (voiced) and whispered conditions. There were three experimental conditions: voiced, whispered, and filtered (255 Hz low-pass filtering of voiced samples). A middle 500 ms segment was excised from each of the sustained vowel samples. A total of 15 listeners judged the speaker sex for each vowel, and also indicated their confidence in their judgments on a seven-point scale. The results showed that 96% were correct for the voiced vowels, 91% were correct for the lowpass filtered vowels, and 75% were correct for the whispered vowels. The confidence ratings correlated well with the correct

identifications. From these results, Lass and his colleague concluded that, in speaker sex identification, $F0$ is a more important acoustic cue than the resonance characteristics of the speaker.

In summary, $F0$ seems to be the most important acoustic cue in speaker sex identification in a normal situation, although somewhat less accurate identification would be possible with vocal tract resonance characteristics alone. It should be emphasized, however, that what has been mentioned above as "vocal tract resonance characteristics" may involve, in a strict sense, both the characteristics of "true" vocal tract resonance *and* those of a glottal (or electrolaryngeal, whispering, or fricative) source except for its fundamental frequency (assuming there is one). In fact, it was revealed that, to achieve natural voice quality in synthesized speech, not only the vocal tract length and fundamental frequency but also the glottal waveform should be specified in accordance with sex distinction (Umeda and Teranishi, 1966). Moreover, there is evidence which shows that the glottal waveform can influence listener judgments of speaker gender (Carrell, 1984), though the effect is small (Mullennix et al., 1991).

In addition, Sato (1974) showed that the slope of the overall spectral envelope, the fundamental frequency, the formant spacings, and the formant bandwidths, should all be taken into account in order to transform the male voice to the female voice. Therefore, there seems to be many perceptual cues available in speaker sex identification.

III. VOICE QUALITY

Voice quality is closely related to (but not the same as) timbre. Timbre has a broader range of meaning than voice quality. Even if we restrict ourselves to the human voice, timbre includes not only voice quality, but also phoneme differences. There are several studies investigating timbre in this sense, in which vowels produced by many speakers were examined (Pols et al., 1969; Klein et al., 1970; Pols et al., 1973). The main conclusion derived in these studies, was that vowel timbre can be represented in a perceptual

space with few dimensions, which is very similar to a spectral space.

When voice quality is the main concern, phoneme differences within each speaker are usually put aside. However, note that differences were observed among vowels as to the contribution of vocal tract characteristics to voice quality (Matsumoto, et al., 1973).

A small number (two to four) of dimensions (or factors) is said to be sufficient to explain the perceptual space of voice similarity (Voiers, 1964; Holmgren, 1967; Matsumoto et al., 1973; Singh and Murry, 1978; Murry and Singh, 1980; Kuwabara and Ohgushi, 1983; Furui and Akagi, 1985; Kreiman et al., 1990).

Most of these studies employed only male speakers (Voiers, 1964; Matsumoto et al., 1973; Kuwabara and Ohgushi, 1983; Furui and Akagi, 1985; Kreiman et al., 1990), although a few studies did investigate the voice quality of both male and female speakers (Singh and Murry, 1978; Murry and Singh, 1980). According to a study which simultaneously covers both male and female voices, the most prominent perceptual dimension is that of male-female (Singh and Murry, 1978).

Most of the studies include physical analysis of the stimuli, and the results of the analysis have been utilized to find, and interpret, the correlation between acoustic cues and psychological dimensions. However, the results of these interpretations are inconsistent: Some investigators emphasized the importance of mean $F0$ (Matsumoto et al., 1973; Murry and Singh, 1980; Furui and Akagi, 1985), while others emphasized the importance of dynamic features (formant pattern, $F0$ pattern) and concluded that mean $F0$ was least important (Kuwabara and Ohgushi, 1983).

There are also other acoustic cues which seem to be relevant to voice quality perception: spectral envelopes (Furui and Akagi, 1985), especially the existence or absence of a peak around 3-4 kHz (Kuwabara and Ohgushi, 1983; Furui and Akagi, 1985), "the slope of glottal source spectrum" and formant frequencies (Matsumoto et al., 1973) and formant ratios (Murry and Singh, 1980). However, it is difficult to determine what acoustic parameter is really important.

One reason may be that, the choice of the physical factors have been made on a rather arbitrary basis. In other words, to begin with, there is no objective (in the strictest sense) criteria in

choosing what kind of physical parameter is to be measured. It is possible to select them by referring to the speech production process, as did most researchers, and this may be the most "economical" way to do that.

However, the parameters actually chosen are quite different from study to study, and the overlap among them is only slight. Thus, it is very hard to compare the results obtained in different studies. Probably the best thing we can do is to measure as many aspects of the physical parameters as possible, and select the ones that the strongly correlated with psychological parameters. Even so, there remains the problem that it is by no means easy to check whether a given psychological effect is caused by a given physical parameter, because some physical parameters correlate with other parameters, and because we have little knowledge about in what form listeners use specific physical information.

Another difficulty in determining significant acoustic parameters is the variability of the strategies utilized by the listeners: they seem to use different cues according to speaker grouping and the stimulus sample (Singh and Murry, 1978; Murry and Singh, 1980), and there is also variability among the listeners themselves as to what cues to use, even for the same pair of speakers (Kuwabara and Ohgushi, 1984; Kreiman et al., 1990). One possible way to solve this problem would be to represent the individual differences as weights on the psychological dimensions, as in the INDSCAL model (Kruskal and Wish, 1978; Schiffman et al., 1981) used in the study of Kreiman et al. (1990).

The other reason may be the insufficient number of listeners, though there would be no definite answer to the question of just how many is sufficient (cf. I. B).

Yet another reason may be in the differences between the speech samples used. Matsumoto et al. (1973) used excised vowels whose durations were 500 ms, and Kreiman et al. (1990) used 1.67-sec-long excised vowels, whereas others used words (Furui and Akagi, 1985), phrases (Singh and Murry, 1978; Kuwabara and Ohgushi, 1983), or sentences (Voiers, 1964; Holmgren, 1967). Since Murry and Singh (1980) found that the similarity judgements of the subjects were influenced by the kind of stimulus (an excised vowel of two seconds and a short phrase), these inconsistencies

may have caused the differences in the results. Furthermore, the group of speakers employed in Kuwabara and Ohgushi (1983) consisted of male announcers and nonprofessional male speakers, i.e., two largely different sub-groups, while in other studies each group of speakers seems to be relatively homogeneous. This may cause another inconsistency.

These considerations lead us to ask, "Is it really meaningful to seek *the* most important physical parameter?" The general answer to this question will be found later in this article.

IV. SPEAKER IDENTIFICATION WITH FAMILIAR VOICES

There would be many kinds of cues for speaker identification in daily life: loudness, pitch, voice quality, idiosyncratic vocabulary, listener expectation, sense and context of a message, and so on. In a laboratory study, however, efforts are usually taken to reduce such factors as speech content differences and loudness, where it seems quite obvious that they can contribute to identification task performance.

Several studies examined the relation between speaker identification and some simple physical parameters, such as the duration of the speech sample (Pollack et al., 1954; Compton, 1963; Bricker and Pruzansky, 1966), the range of $F0$ (Compton, 1963), and the mean $F0$ (Abberton and Fourcin, 1978; Carrell, 1984). It has been revealed that the duration *per se* is not very important (Pollack et al., 1954) but that the number of phonemes does affect the performance (Bricker and Pruzansky, 1966), and that the $F0$ range and the mean $F0$ both influence speaker identification (Compton, 1963; Abberton and Fourcin, 1978; Carrell, 1984).

Recently, the role of acoustic pattern or contour in speaker identification has been emphasized. The importance of acoustic pattern in speaker identification was first revealed by the experiments of Bricker and Pruzansky (1966). They found that all reversed listening scores were below the forward scores, but well above chance. Since then, it has been shown in a study using laryngographic techniques that the $F0$ contour provides important speaker identification information (Abberton and Fourcin, 1978).

Furthermore, based on a study that employed a fairly large number of subjects (45 famous speakers and 94 listeners), it has been argued that each familiar voice should be viewed as a relatively unique pattern, since the degree of the effects of backward presentation on voice recognition (their tasks were to "recognize" the names) differed from voice to voice (Van Lancker, Kreiman, and Emmorey, 1985). Thus Van Lancker and her colleagues claimed:

> "...it is not useful to pursue *the* parameter that contributes most universally to voice identity. Instead, many parameters and combinations of parameters constitute a pool from which certain selected cues are utilized for recognition."

The same conclusion is supported by the fact that some voices were easily recognized even when the speaking rates were altered, while others were not (Van Lancker, Kreiman, and Wickens, 1985).

Thus, it is now clear that no single acoustic cue is universally important for speaker identification. What cues to use and how to use them vary from situation to situation. In other words, we can say that a certain parameter may contribute to speaker identification, but cannot say, whatever it is, that it is the most important.

V. DISCUSSION

The main conclusion of this article could be put as follows. Each voice is to be viewed as a unique pattern which contains many different acoustic parameters, and listeners utilize different parameters for different situations. Accordingly it is almost meaningless to try to determine the most important parameter without specifying the situation.

This conclusion might, at first sight, appear to conflict with the conclusion reached from sex identification studies, which can be summarized as showing that $F0$ is most important. However, this can be explained by the fact that in the sex identification task, the stimuli are from, in effect, two largely different groups (male and

female), while in the speaker identification and voice quality judgment task, the differences among the stimuli are much subtler. That is, if the differences are large, subjects will be apt to use the most prominent cue to perform the task but, if not, there would be little difference in the prominence of the cues, and subjects will use any available cue to perform the task.

To show the importance of the temporal pattern, or to eliminate that cue, backward presentation has been utilized (Bricker and Pruzansky, 1966; Coleman, 1973; Van Lancker, Kreiman, and Emmorey, 1985). However, this presentation method not only alters the temporal pattern of the stimulus but may also alter the timbre, even if the average spectrum is the same (cf. Ueda and Akagi, 1990). Thus, this technique should be applied carefully if voice quality is the main concern.

Except for excised vowels, the initial part of a stimulus may contain a cue for the judgment of voice quality: idiosyncratic patterns of articulation or voicing may result in a specific acoustic pattern. If we compare the results using stimuli containing their initial parts with stimuli not containing their initial parts, we may be able to show the importance of the information involved in the initial parts of the stimuli.

However, this theme itself does not seem overly promising, since studies on instrumental sounds have already revealed that the initial transients or amplitude envelopes of the stimuli are very important for timbre perception and instrument identification (Berger, 1964; Saldanha and Corso, 1964; Strong and Clark, 1967; Wedin and Goude, 1972; Miller and Carterette, 1975; Grey, 1977; Ando and Yamaguchi, 1983). As a speech perception research project, this theme would have some meaning only when we can trace the process from the articulatory gesture, through the specific acoustic pattern, to the perception. There is no need to think that the auditory system *per se* is somewhat special in perceiving transients produced by the speech organs.

ACKNOWLEDGMENT

Thanks are due to Dr. Noriko Kobayashi for discussion, and Dr. Yoh'ichi Tohkura for his support on this research.

Abberton, E., and Fourcin, A. J. (1978). "Intonation and speaker identification," Language and Speech, 21, 305-318.

Ando, S., and Yamaguchi, K. (1983). "Considerations on physical characteristics of samisen tones," J. Acoust. Soc. Jpn. 39, 433-443 (in Japanese).

Bennett, S., and Weinberg, B. (1979). "Acoustic correlates of perceived sexual identify in preadolescent children's voices," J. Acoust. Soc. Am. 66, 989-1000.

Berger, K. W. (1964). "Some factors in the recognition of timbre," J. Acoust. Soc. Am. 36, 1888-1891.

Bricker, P. D., and Pruzansky, S. (1966). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. 40, 1441-1449.

Carrell, T. D. (1984). "Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification," Research on Speech Perception, Technical Report No. 5 (Department of Psychology, Indiana University, Bloomington).

Coleman, R. O. (1971). "Male and female voice quality and its relationship to vowel formant frequencies," J. Speech Hear. Res. 14, 565-577.

Coleman, R. O. (1973). "Speaker identification in the absence of inter-subject differences in glottal source characteristics," J. Acoust. Soc. Am. 53, 1741-1743.

Coleman, R. O. (1976). "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice," J. Speech Hear. Res. 19, 168-180.

Compton, A. J. (1963). "Effects of filtering and vocal duration upon the identification of speakers, aurally," J. Acoust. Soc. Am. 35, 1748-1752.

Furui, S., and Akagi, M. (1985). "Perception of voice individuality and physical correlates," Trans. Committee of Hearing Res. H-85-18 (in Japanese).

Grey, J. M. (1977). "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. 61, 1270-1277.

Holmgren, G. L. (1967). "Physical and psychological correlates of speaker recognition," J. Speech Hear. Res. 10, 57-66.

Ingemann, F. (1968). "Identification of the speaker's sex from voiceless fricatives," J. Acoust. Soc. Am. 44, 1142-1144.

Itoh, K., and Saito, S. (1986). "Effects of acoustical speech feature parameters on perceptual identification of speaker," Report of the E. C. L., NTT, Jpn. 35, 677-686 (in Japanese).

Iwashita, T. (1979). *Semantics of Osgood and SD method* (Kawashima, Tokyo) (in Japanese).

Klein, W., Plomp, R., and Pols, L. C. W. (1970). "Vowel spectra, vowel spaces, and vowel identification," J. Acoust. Soc. Am. 48, 999-1009.

Kobayashi, N. (1991). Personal communication.

Kreiman, J., Gerratt, B. R., and Precoda, K. (1990). "Listener experience and perception of voice quality," J. Speech and Hearing Res. 33, 103-115.

Kruskal, J. B., and Wish, M. (1978). *Multidimensional Scaling* (Sage, London).

Kuwabara, H., and Ohgushi, K. (1983). "Acoustic characteristics of professional male announcers speech," IECE, J66-A, 545-552 (in Japanese).

Kuwabara, H., and Ohgushi, K. (1984). "Experiments on voice qualities of vowels in males and females and correlation with acoustic features," Language and Speech, 27, 135-145.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., and Bourne, V. T. (1976). "Speaker sex identification from voiced, whispered, and filtered isolated vowels," J. Acoust. Soc. Am. 59, 675-678.

Matsumoto, H., Hiki, S., Sone, T., Nimura, T. (1973). "Multidimensional representation of personal quality of vowels and its acoustical correlates," IEEE Trans. AU-21, 428-436.

Miller, J. R. and Carterette, E. C. (1975). "Perceptual space for musical structures," J. Acoust. Soc. Am. 58, 711-720.

Mullennix, J., Johnson, K., and Topcu, M. (1991). "Context effects in the perception of personal information in the speech signal," J. Acoust. Soc. Am. 89, 2011, 9SP11.

Murry, T., and Singh, S. (1980). "Multidimensional analysis of male and female voices," J. Acoust. Soc. Am., 68, 1294-1300.

Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175-184.

Pollack, I., Pickett, J. M., and Sumby, W. H. (1954). "On the identification of speakers by voice," J. Acoust. Soc. Am. 26, 403-406.

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

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

Saldanha, E. L. and Corso, J. F. (1964). "Timbre cues and the identification of musical instruments," J. Acoust. Soc. Am. 36, 2021-2026.

Sato, H. (1974). "Acoustic cues of female voice quality," Electronics and communications in Japan, 57-A, 29-38.

Schiffman, S. S., Reynolds, M. L., and Young, F. W. (1981). *Introduction to Multidimensional Scaling* (Academic, Orlando).

Schwartz, M. F. (1968). "Identification of speaker sex from isolated, voiceless fricatives," J. Acoust. Soc. Am. 43, 1178-1179.

Schwartz, M. F., and Rine, H. E. (1968). "Identification of speaker sex from isolated, whispered vowels," J. Acoust. Soc. Am. 44, 1736-1737.

Shearme, J. N., and Holmes, J. N. (1959). "An experiment concerning the recognition of voices," Language and Speech, 2, 123-131.

Singh, S., and Murry, T. (1978). "Multidimensional classification of normal voice qualities," J. Acoust. Soc. Am. 64, 81-87.

Strong, W. and Clark, M. Jr. (1967). "Perturbations of synthetic orchestral wind-instrument tones," J. Acoust. Soc. Am. 41, 277-285.

Ueda, K., and Akagi, M. (1990). "Sharpness and amplitude envelopes of broadband noise," J. Acoust. Soc. Am. 87, 814-819.

Umeda, N., and Teranishi, R. (1966). "Phonemic feature and vocal feature--Synthesis of speech sounds, using an acoustic model of vocal tract," J. Acoust. Soc. Jpn. 22, 195-203 (in Japanese).

Van Lancker, D., and Kreiman, J. (1987). "Voice discrimination and recognition are separate abilities," Neuropsychologia, 25, 829-834.

Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). "Familiar voice recognition: patterns and parameters. Part I: Recognition of backward voices," J. Phonetics, 13, 19-38.

Van Lancker, D., Kreiman, J., and Wickens, T. D. (1985). "Familiar voice recognition: patterns and parameters. Part II: Recognition of rate-altered voices," J. Phonetics, 13, 39-52.

Voiers, W. D. (1964). "Perceptual bases of speaker identity," J. Acoust. Soc. Am. 36, 403-406.

Wedin, L. and Goude, G. (1972). "Dimension analysis of the perception of instrumental timbre," Scand. J. Psychol. 13, 228-240.