Random Splicing: A Method of Investigating the Effects of Voice Quality on Impression Formation

Mihoko Teshigawara

Department of Linguistics
University of Victoria, Canada
mteshi@myrealbox.com

Abstract

This paper discusses an experiment in which 32 subjects listened to random-spliced excerpts from the speech of 27 cartoon characters and rated their impressions of age, gender, physical and personality traits, emotional states, and vocal characteristics. Statistical analyses were performed in order to examine the consistency of participants' trait ratings and the relationship between the auditory characteristics of the voices and subjects' trait ratings of the speakers. The results suggest that random splicing can be a useful method of examining the effects of voice quality on impression formation.

1. Introduction

Previous studies on vocal stereotypes reveal that people infer similar personality traits from voices [1]. Of the vocal stereotype studies, most have been interested in the effects of prosody (i.e., pitch, loudness, and rate) on impression formation, which have been investigated using computer programs or systematic control by speakers [2, 3]. However, to my knowledge, only one study has examined the effects of voice quality on impression formation systematically [4]. In a similar vein, even though there appears to be ever-increasing interest in research on vocal cues to emotion and to paralinguistic information, most of the studies are concerned with the acoustic correlates of prosody, and only a few have considered voice quality [5, 6, 7]. (See [8] for a detailed review of the studies along these lines.) This study attempts to fill these gaps in our knowledge by investigating the phonetic correlates of vocal stereotypes in an experimental setting using Japanese cartoon (*anime*) voices – samples that are considered to reflect the physical attributes and personality traits of characters and the vocal stereotypes that consumers, filmmakers, and voice actors share.

In order to elicit listeners' responses to the voices independent of verbal content, it is necessary to control the contents of the speech samples by using standardized speech materials or by masking the contents. Due to the nature of the present study's corpus, which consists of 20 different animated cartoons (see [8] for more details), only the latter was possible in this study. The following content-masking techniques, along with their effects on expert ratings and laypersons' perceptions have been proposed and studied: low-pass filtering, random splicing, backward speech, pitch inversion, tone-silence sequences, and reiterant speech [9, 10, 11]. According to Scherer et al. [10] and van Bezooijen and Beves [11], of these techniques (with the exception of reiterant speech, which was investigated by Friend and Farrar, [9]), random splicing is the only one that retains voice quality information, which is the focus of the present study. In random splicing, speech samples are divided into small segments (250 ms is conventional) and rearranged in an order different from the original. In van Bezooijen and Beves' study, where the effects of random splicing and low-pass filtering on expert ratings of voice quality and prosodic settings were investigated, random splicing was found to retain information pertaining not only to voice quality features such as harshness, denasality, and pharyngeal constriction, but also to prosodic features such as pitch level and loudness [11]. Although Friend and Farrar [9] do not discuss the effects of reiterant speech, this technique, which involves replacing the syllables of the original speech with three meaningless syllables ([ba], [ma], or [sa]), appears to reduce a significant amount of voice quality information. Therefore, to investigate the effects of different voice quality settings, random splicing appears to be the best technique and was adopted in the stimuli preparation.

It should be noted, however, that it has been suggested that the random-splicing technique may introduce systematic biases to perception. Examining the effects of random splicing and four other masking techniques on deception detection, Scherer et al. found that impressions of "relaxed" were rated higher in the dishonest condition than in the honest condition, while other personality traits were rated comparably with the original forward playing mode – that is, in the latter mode, the honest condition was rated higher than the dishonest condition [10]. Scherer et al. speculate that the perception of "relaxed" may depend on factors such as pausing and tempo, which are lost in random splicing. Friend and Farrar studied the effects of low-pass filtering, random splicing, and reiterant speech on judgments of the speaker's affective states; they found that random splicing increased ratings of anger and excitement [9]. Thus, it is necessary to interpret the results of the present study with caution, especially personality, emotion, and vocal trait items related to those mentioned above.

In this study, 27 *anime* characters' voices were selected based on the auditory characteristics of voices of heroes and villains identified in a separate study [8]. Thirty-two subjects listened to random-spliced speech excerpts from the 27 target cartoon characters and rated their impressions of the speakers using trait items in the following four categories: physical traits, personality traits, emotional states, and vocal characteristics.

2. Method

2.1. Stimuli

Prior to the present paper, the author identified auditory characteristics of the voices of 88 characters (44 heroes, 42 villains, and two supporting roles) from 20 animated cartoons, using Laver's framework for voice quality description [12,

13]. The following summarizes the auditory characteristics identified across categories (see [8] for more details):

1. Heroes' voices exhibited an absence of laryngeal constriction and the presence of breathy voice.
2. The majority of villains' voices exhibited laryngeal constriction (including larynx raising) and harsh voice caused by tense laryngeal tension settings; however, larynx lowering was observed in a majority of female and some male villains. (Laryngeal constriction is that of the "well-understood aryepiglottic sphincter mechanism" [14].)

In order to investigate whether the identified auditory characteristics contribute to people's perceptions of good and bad characters, Japanese laypersons' perceptions of selected speech samples were examined in an experimental setting. It was hypothesized that participants would attribute less favorable physical traits, personality traits, emotional states, and vocal characteristics to speakers who exhibited laryngeal constriction/larynx lowering no matter which roles they played in the original cartoons.

In light of the auditory characteristics summarized above, the 88 character voices were divided into two groups, representative and non-representative: representative meaning that characters exhibited auditory characteristics appropriate to their role, and non-representative meaning that characters exhibited auditory characteristics opposite to or simply atypical of their role. Within these two groups, characters were examined according to role, gender and age (adult vs. child). For example, villains showing either laryngeal constriction or larynx lowering were categorized into the representative villain group, while those showing neither trait fell into the non-representative villain group. There were 16 possible groups: hero or villain (2) × gender (2) × age (2) × representativeness (2). However, since there was only one child villain (male) in the corpus, this classification system yielded only 13 groups. Two speakers were chosen for each of 12 groups, with the exception of the child male villain group, which had only one speaker. In addition, two supporting roles (one male child and one female adult) exhibiting the auditory characteristics of villains' voices, that is, laryngeal constriction and harsh voice, were added in order to see whether they would be rated similarly to heroes or villains. Therefore, the voices of 27 speakers in total were chosen as the basis for experimental stimuli.

Noise-free speech samples of these 27 speakers had been stored on a personal computer for auditory and acoustic analyses for a separate study [8]. They had been recorded from VHS tapes of the cartoons. First, in order to create stimuli representative of each speaker, speech portions produced with a voice quality setting deviating from the speaker's normal setting were removed, with the exception of characters who were consistently angry or shouting. Intensities were standardized across speakers so that the maximum intensity was between 70 and 72 dB. Following previous research using the random-splicing technique ([9, 10, 11], after removing pauses, the digitized speech samples were divided into 250-ms segments. The first and last 3 ms of each segment were linearly attenuated to zero amplitude in order to avoid the introduction of transients [9]. In order to create a 5-s stimulus for each speaker, 20 250-ms segments were prepared and rearranged so that segments could not occur in the same relative order in the spliced stimulus as in the original.

In order to counterbalance the effects of ordering, two stimulus orders (A and B) were used: in A, the 27 speakers were randomly ordered disregarding the speaker groups, while B was the reverse of A. For each speaker, the speaker number was announced followed by the 5-s stimulus; after 1 s of silence, the same segment was repeated, followed by 70 s of silence. This gave participants a total of 81 s to rate each speaker, which, according to previous studies (e.g., [15]), is considered sufficient to rate the 21 trait items selected in this experiment. Participants were given a practice session in which they rated an additional three speakers before rating the 27 target speakers.

2.2. Procedures and Participants

Twenty-one trait items were chosen to be used in the questionnaire for the rating session. English translations are given for the items as follows: gender (female or male); age group (0–10; 11–18; 19–35; 36–60; over 61); physical characteristics ("big," "good-looking"); personality traits, 11 in total, of which three were chosen for their pertinence to heroes of Japanese *anime* in particular ("selfless," "loyal," "devoted") [16], three were thought to be universal characteristics of heroes ("brave," "intelligent," "strong"), and five represented each of the five factors in the NEO Personality Inventory ("sociable," "calm," "curious," "conscientious," "sympathetic") [17]; emotional states ("positive emotion"); and vocal characteristics ("high-pitched," "loud," "relaxed," "pleasant," "attractive"). The 19 adjectives were rated on 7-point scales, from 1 (*not at all true*) to 7 (*extremely true*). In the questionnaire, in order to counterbalance ordering effects, the order of the six trait categories and, where applicable, the items within trait categories were systematically varied, yielding four questionnaire types (Ia, Ib, IIa, IIb) (see [8] for more details).

Thirty-two participants (15 males, 17 females; average age 22.8 years old) were recruited from Nagoya University, Japan and the vicinity. In total, eight experimental conditions were yielded, combining the two stimulus orders (A and B) and the four questionnaire types. Four participants were assigned to each experimental condition, with the exception of the two groups that used Questionnaire Ia, in which five participants listened to stimulus order A and three listened to B.

Experimental sessions were run in groups of up to seven in a soundproof room in the School of Letters building at Nagoya University. Using a CD player, the experimenter played a CD containing instructions recorded by the author, a practice session and the 27 target stimuli. The same instructions were given in the questionnaire booklet as well. Participants were told that they would hear two 5-s content-masked cartoon speech excerpts for each speaker, and they were asked to rate impressions of the speakers' traits on 7-point scales and choose appropriate groups for gender and age. After the experiment, participants completed a questionnaire including demographic information about themselves and their exposure to *anime*. Each session lasted less than one hour.

3. Results and Discussion

In order to examine the consistency of participants' trait ratings, Cronbach's alpha was calculated for each of the 21 trait items across participants. SPSS version 11.5 was used for the statistical analyses performed in this paper. The Cronbach's alphas were very high, ranging between .90 and .99 for all but two items (.87 for "sociable" and .80 for

"positive emotion"). This was the case with the two prosodic items, namely, "high-pitched" and "loud," as well (0.98 and 0.92, respectively), which is in agreement with the results from a study using expert raters [11]. These results suggest that random splicing can be a reliable technique for examining the effects of voice quality on impression formation. The high Cronbach's alphas are comparable to the values obtained in studies using stimuli that were not random-spliced [15]. After a close inspection of the raw data and additional bivariate correlations calculated for each pair of participants and for each participant relative to the average ratings of all participants, it was decided that results for one participant would be removed from the rest of the analysis.

In a separate study, a series of repeated measures ANOVAs was carried out for each of the selected 16 items in order to examine whether participants responded to the stimuli according to the differences in epilaryngeal state among the characters [8]. In addition to significant main effects for the factor role (i.e., hero vs. villain), significant interaction between role and representativeness emerged in 12 out of 16 items, suggesting that participants attributed less favorable physical traits, personality traits, emotional states, and vocal characteristics to speakers who exhibited non-neutral epilaryngeal states (i.e., laryngeal constriction or more than a slight degree of larynx lowering) regardless of the roles they played in the original cartoons. Therefore, it can be said that the classification of auditory characteristics into representative and non-representative based on the epilaryngeal states identified in [8] was valid.

Next, correlations were calculated for the 19 adjective trait items for all 27 speakers used in the experiment, averaging the points selected by the participants for each speaker. There were 95 significant correlations of 171 possible combinations of traits (55.6% of possible combinations). In order to better understand relationships among the adjective items, a factor analysis was performed using the mean ratings for the 19 items for each speaker. Principal components analysis with iteration was used, followed by varimax rotation. According to the scree test, it was decided that the first five factors would be kept. The five factors accounted for 95.2% of the total variance (see Table 1).

Factor 1 consists of six items, three of which were chosen because they were thought to represent the characteristics of Japanese *anime* heroes ("selfless," "devoted," and "loyal"), followed by the three factors from the NEO Personality Inventory, "conscientious," "sympathetic," and "sociable," representing Conscientiousness, Agreeableness, and Extraversion, respectively, which were expected to be independent of one other [17]. Because the first three traits were particularly relevant to Japanese *anime* heroes, this factor was named Heroicness. Factor 2 consists of six items: one physical characteristic, namely, "good-looking"; three vocal characteristics, namely, "attractive," "pleasant," and "loud," with the last item reversed; and two personality traits, namely, "intelligent" and "calm." In a separate study, it was found that the items constituting Factor 2 had strong correlations with degrees of laryngeal constriction in the voice [8]. (The items constituting Factor 1 also had moderate to strong negative correlations with the degree of laryngeal constriction [8].) However, in the naming of this factor, the term Desirability was adopted rather than laryngeal constriction, in recognition of the desirable qualities these items share. Factor 3 again consists of items from three different categories: "strong" and "brave" from personality traits, "big" from physical characteristics, and "high-pitched" from vocal characteristics, which is reversed. While attributes such as "strong" and "brave" were included among the personality trait items, it is likely that participants associated them with physical strength. Factor 4 consists of two items: one is "positive emotion" and the other is "relaxed" from the vocal characteristics. This factor was named Emotional Stability. Factor 5 consists of only one item, namely, "curious"; because of the origin of this item (i.e., Openness from the NEO Personality Inventory), it was named Openness. To sum up, it can be said that the five factors that emerged in this analysis seem to be reasonable, and that Factors 1 and 2 are interpretable in terms of the degrees of laryngeal constriction in the voice, which is a voice quality setting.

Table 1: *Rotated factor matrix for perceptual experiment adjective items*

Adjectives (category)	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Selfless (p.)	**0.89**	0.39	-0.04	-0.02	-0.14
Devoted (p.)	**0.86**	0.02	-0.31	-0.09	0.34
Loyal (p.)	**0.86**	0.46	0.11	0.15	0.00
Consc. (p.)	**0.85**	0.38	0.19	0.02	-0.23
Symp. (p.)	**0.79**	0.53	-0.13	0.19	0.13
Sociable (p.)	**0.67**	0.18	-0.39	0.14	0.49
G.-l. (ph.)	0.42	**0.89**	0.04	-0.03	0.08
Attract. (v.)	0.41	**0.86**	0.14	0.19	0.11
Pleasant (v.)	0.36	**0.83**	0.16	0.36	0.09
Intelli. (p.)	0.27	**0.77**	0.40	0.04	-0.32
Calm (p.)	0.40	**0.72**	0.17	0.42	-0.32
Loud (v.)	-0.12	**-0.71**	0.09	-0.33	0.49
Strong (p.)	-0.10	0.21	**0.94**	-0.08	0.03
Big (ph.)	0.02	0.12	**0.92**	-0.13	-0.21
High (v.)	0.30	0.12	**-0.87**	-0.16	0.27
Brave (p.)	0.48	0.25	**0.73**	0.23	0.24
Positive (e.)	0.15	0.13	-0.16	**0.93**	0.13
Relaxed (v.)	-0.14	0.49	0.19	**0.82**	-0.03
Curious (p.)	0.12	-0.20	-0.53	0.37	**0.70**
Eigenvalue	8.96	4.79	2.25	1.45	0.65
Cumulative explained variance (%)	47.14	72.35	84.21	91.82	95.22

Note. E., p., ph., and v. stand for emotional state, personality traits, physical characteristics, and vocal characteristics, respectively.

However, there is also a possibility that prosody, notably pitch, may have played a role in the participants' ratings. In [8], it was revealed that the items constituting Factor 4, i.e., "positive emotion" and "relaxed," were generally rated lower than the other items: the average ratings for these two traits across participants were 3.68 for heroes and 3.21 for villains, while the averages for the remaining positive traits (14 adjectives excluding "big," "high-pitched," and "loud") were 4.50 for heroes and 3.83 for villains. It is possible that these results are an artifact of the random splicing technique, in other words, that the effects of the epilaryngeal states may not have been genuine; however, this does not seem to be the case. (See Section 5.2.8 of [8] for more details for each speaker's combined rating for the items concerned, epilaryngeal state and mean F0s.) When the 27 speakers were ordered according to the combined ratings for the items concerned, the rough descending order was as follows: (i) a majority of heroic

voices (i.e., representative heroes, non-representative heroes and villains); (ii) villainous voices with laryngeal constriction; (iii) a minority of heroic voices; and (iv) villainous voices with larynx lowering. (Heroic voices in (i) and (iii) contain those with slight or intermittent laryngeal constriction and those with slight to moderate larynx lowering.) Table 2 summarizes the breakdown of the four groups and the combined mean rating (out of 14 points) averaged across speakers for each group.

Considering that voices exhibiting non-neutral epilaryngeal states were rated unfavorably for other positive trait items in general, it makes sense that the villainous voice groups (ii) and (iv) received lower ratings than the heroic voice group (i) for these items as well. However, it was not expected that the minority of heroic voices (iii) would receive lower ratings than a villainous voice group (ii). A close examination of the raw data revealed that the difference between (i) and (iii) was F0 – most of the (iii) voices had very high F0: for the characters played by male voice actors, the mean F0 across speakers in (i) was 157.0 Hz, as compared with 285.6 Hz for (iii); for those played by female voice actors, the mean F0 was 366.2 Hz for (i), and 432.5 Hz for (iii). Therefore, it is possible that the participants' ratings may have been influenced by the unusually high F0 of the speakers in (iii). These results may suggest that the voices with deviant characteristics in voice quality and/or pitch could have contributed to the negative ratings for these items and that the trait ratings obtained using random splicing may reflect some prosodic information as well as voice quality. However, in order to determine the true source of these differences, that is, which effects can be attributed to epilaryngeal states (voice quality) and pitch height, it would be necessary to compare the present results with those from experiments using stimuli that are not random-spliced.

4. Conclusions

This paper reported on the results of an experiment in which 32 subjects listened to random-spliced speech excerpts from 27 cartoon characters and rated their impressions of age, gender, physical and personality traits, emotional states, and vocal characteristics. Based on the high reliabilities and on the two factors (i.e., Factors 1 and 2) that are interpretable in terms of a voice quality setting, the findings suggested that random splicing can indeed be a useful method of examining the effects of voice quality on impression formation. However, it was also noted that the resulting random-spliced stimuli could retain some prosodic information, such as mean pitch height, as well.

5. References

[1] Zuckerman, M.; Miyake, K., 1993. The attractive voice: What makes it so? *Journal of Nonverbal Behavior* 17, 119–135.

[2] Nass, C.; Lee, K.M., 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7, 171–181.

[3] Ray, G.B., 1986. Vocally cued personality prototypes: An implicit personality theory approach. *Communication Monographs* 53, 266–276.

[4] Addington, D.W., 1968. The relationship of selected vocal characteristics to personality perception. *Speech Monographs* 35, 492–503.

[5] Campbell, N.; Mokhtari, P., 2003. Voice quality: The 4th prosodic dimension. *Proceedings of the 15th International Congress of Phonetic Sciences.* Spain: Universitat Autònoma de Barcelona, 2417–2420.

[6] Fujimoto, M.; Maekawa, K., 2003. Variation in phonation types due to paralinguistic information: An analysis of high-speed video images. *Proceedings of the 15th International Congress of Phonetic Sciences.* Spain: Universitat Autònoma de Barcelona, 2401–2404.

[7] Gobl, C.; Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40, 189–212.

[8] Teshigawara, M., 2003. Voices in Japanese Animation: A Phonetic Study of Vocal Stereotypes of Heroes and Villains in Japanese Culture. Unpublished PhD dissertation, University of Victoria, Canada. [available from http://web.uvic.ca/ling/graduate/theses-dissertations.htm]

[9] Friend, M.; Farrar, M.J., 1994. A comparison of content-masking procedures for obtaining judgments of discrete affective states. *Journal of the Acoustical Society of America* 96, 1283–1290.

[10] Scherer, K.R.; Feldstein, S.; Bond, R.N.; Rosenthal, R., 1985. Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research* 14, 409–425.

[11] van Bezooijen, R.; Boves, L., 1986. The effects of low-pass filtering and random splicing on the perception of speech. *Journal of Psycholinguistic Research* 15, 403–417.

[12] Laver, J., 1994. *Principles of Phonetics.* Cambridge, UK: Cambridge University Press.

[13] Laver, J., 2000. Phonetic evaluation of voice quality. In *Voice Quality Measurement*, R.D. Kent & M.J. Ball (eds.). San Diego, CA: Singular, 37–48.

[14] Esling, J. H., 1999. The IPA categories "pharyngeal" and "epiglottal": Laryngoscopic observations of pharyngeal articulations and larynx height. *Language and Speech* 42, 349–372.

[15] van Bezooijen, R., 1995. Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech* 38, 253–265.

[16] Levi, A., 1998. The new American hero: Made in Japan. In *The Soul of Popular Culture: Looking at Contemporary Heroes, Myths, and Monsters*, M.L. Kittelson (ed.). Chicago: Open Court, 68-83.

[17] McCrae, R.R.; Costa, P.T., Jr., 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology* 52, 81–90.

Table 2: *Order of four voice groups according to combined ratings for "positive emotion" and "relaxed"*

Voice group	Number of speakers	Combined mean rating (out of 14)
(i) Heroic	13	7.86
(ii) Villainous (constr.)	5	6.66
(iii) Heroic	7	5.99
(iv) Villainous (lowering)	2	4.90

Note. For voice groups, see text.