A Perception Study on Voice Quality and Stance in Mandarin Chinese

Qiuyuan Li and Peggy Mok

Chinese University of Hong Kong
qiuyuan.li@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

ABSTRACT

The social meaning of voice quality has been gaining more attention in sociophonetics in recent years. However, voice quality and its social significance in Mandarin Chinese are still under-researched. This study conducts a perceptual experiment to explore the role of voice quality in stancetaking in Mandarin Chinese. The results show that listeners can make different value judgments about an object based on the different voice qualities used by the speaker to describe it and construct different portraits of the speaker. In Mandarin Chinese, creaky voice leads listeners to place lower values on the object described and to perceive speakers as more gender-neutral and less good-looking. The findings of this study may help investigate the role of voice quality from the production side and the construction of identities and expressing broader social meanings in Mandarin Chinese in the future.

Keywords: voice quality, Mandarin, stance, sociophonetics.

1. INTRODUCTION

Voice quality, by its narrower definition, refers to the status of vocal fold vibration in voiced segments. Voice quality is widely used to contrast meaning in some African languages. However, in several of the most popular world languages, such as English, Spanish, and Mandarin Chinese, voice quality does not function to contrast lexical meaning but rather to express specific social meanings. For example, in American English, young women who use creaky voice are perceived as valley girls, e.g. [1], while they are also perceived as more urban and more educated. In Japanese, [2] shows that Japanese female voice actors use a specific "sweet voice" quality to express the sweetness of the character's personality. However, although some previous studies have examined the social significance of voice quality in Mandarin, e.g. [3], which has found that the evaluation of creaky voice in Mandarin is associated with its prosodic context (IP-final creaky stimuli are rated as more enthusiastic and interested than other positions), it is insufficient in terms of both quantity and depth compared to the research on it in Western languages. Sociolinguistics has developed its third wave of theorizing in variation study ([4]), which shifts the focus of research from macrosocial categories and ethnic groups to individual speakers and subtle identity constructions, yet sociolinguistic research in Chinese is far behind, for example, no Chinese studies have focused on the role of phonetic features in stancetaking for the time being. Though stance has not been given a universal definition in linguistic research, and it may vary from the point of view that speakers want to express ([5]) to the identity they want to construct ([6]) using certain linguistic markers, the stance investigated in this study refers to the perception listeners have of the speakers and the object described by the speaker based on certain linguistic markers. In this study, participants' stance towards the narrator is revealed by their selection of the AI-generated faces, and their stance towards the object described in the narrative passage is indicated by their responses to the 3 statements about the cake.

Hence, this study investigates whether participants' backgrounds affect how they portray the speaker with different voice qualities and whether non-modal voice qualities may be interpreted to express different stances.

2. METHOD

An online experiment was conducted in which participants reported their backgrounds and listened to a short passage spoken in one of the three voice qualities (breathy, modal, and creaky), then rated against the object described by the passage and chose an AI-generated face that they thought would fit the voice the best. Details of the experimental design are described below.

2.1. Experiment design

The online experiment includes four stages, as shown below in Figure 1.

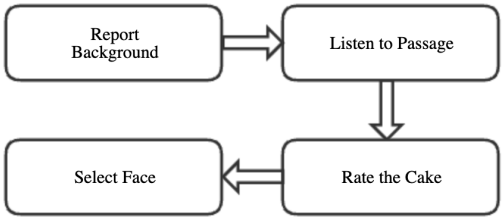

Figure 1: 4 stages in the experiment.

In the first stage, participants reported their backgrounds: gender, age, language spoken at home (Mandarin or not), and Mandarin proficiency.

Participants were then asked to listen to a pre-recorded short passage describing a strawberry cake in the second stage. They were told that a friend bought the cake as a birthday present for them, and the narrator of this passage was exactly that friend. The speaker's gender and the voice quality of this passage were randomized to be female or male, and breathy, modal, or creaky, respectively, while the content remained the same (information on speakers and their voices are to be given in later sections). Participants could listen to the passage as many times as they wanted until they felt comfortable moving on to the rating questions.

The third stage consisted of three statements about the cake described by the passage:

(i). The cake is quite delicious.

(ii). The cake is quite expensive.

(iii). The cake is too much as a gift from this friend and makes you stressed.

The first statement on deliciousness was about the internal evaluation that participants make towards the cake, while the second statement concerned an external and objective value, namely price. The third statement aimed to elicit participants' judgments on the relationship with the narrator.

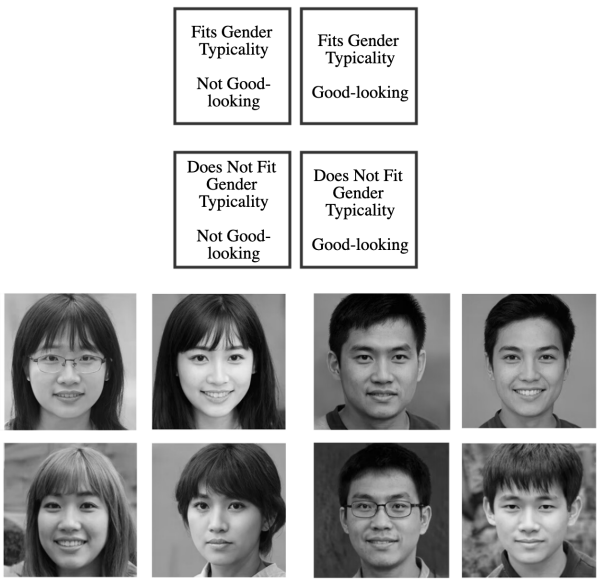

Figure 2: Arrangement of the 4 faces (top) and the actual faces used (female voice bottom-left; male voice bottom-right). Details are explained in section 2.2.

The rating questions were in the format of Likert scale from 1 to 5, with 1 being "strongly disagree" and 5 being "strongly agree". Participants needed to select from 1 to 5 for each statement. This section was designed to investigate the role of voice quality on the perceived stance of the narrator towards the cake.

After rating each statement, participants were then asked to select 1 out of 4 AI-generated faces (details above) that they thought fit the impression of this "friend" the most.

As shown in Figure 2, the 4 faces were arranged in a 4-quadrant, representing a combination of gender typicality and degree of good-looking. The arrangement was fixed in this order. This section was designed to investigate the role of voice quality on participants' impressions of the speaker.

2.2. Stimuli

This experiment contained audio stimuli and visual stimuli.

There were 6 audio stimuli recorded for this experiment: 3 from a female speaker and 3 from a male speaker.

Both speakers were in their mid-20s and spoke Mandarin natively. The female speaker was a research assistant in the phonetics lab, and the male speaker was an amateur singer. They both have good control over the production of different voice qualities.

A few modifications were made to the audio recordings using a Praat plugin, Praat Vocal Toolkit ([7]). Firstly, the loudness of all sound files was adjusted to 60dB. Secondly, for the female recordings, the fundamental frequency median was adjusted to 190Hz, and that of male recordings was 110Hz. Finally, after noticing that the breathy and modal versions of female recordings have a larger pitch variation and might sound unnaturally dramatic, the pitch variation of these two recordings was reduced to 80% of the original, as it gave the most natural sounding intonation according to 4 naïve listeners.

Gender	Voice Quality	F0 Median	Pitch Variation	Loudness
F	Breathy	190Hz	80%	60dB
F	Modal	190Hz	80%	60dB
F	Creaky	190Hz	100%	60dB
M	Breathy	110Hz	100%	60dB
M	Modal	110Hz	100%	60dB
M	Creaky	110Hz	100%	60dB

Table 1: Parameters of processed audio stimuli.

The length of pauses is also modified to keep all recordings at roughly the same speech rate.

Quality	Gender	H1A1	CPP	HNR05
Breathy	F	14.65	14.98	24.75
Breathy	M	31.56	17.62	18.03
Modal	F	13.65	17.94	33.31
Modal	M	22.08	24.95	25.61
Creaky	F	12.66	19.34	22.07
Creaky	M	25.11	23.29	14.8

Table 2: Voice quality acoustics of stimuli.

We have also measured the voice quality acoustics of the stimuli and ensured that their patterns are typical to specific voice qualities as summarized in [8], as shown in Table 2.

The visual stimuli are human faces generated by [9], a group of face generators based on the Neuro-Network algorithm styleGAN2. The author of this generator provides tens of thousands of free and open-source faces generated by the algorithm, and the stimuli used in this experiment are selected from the dataset.

The stimuli have passed 2 rounds of selection. In the first round, about 50 faces were selected from the raw dataset of [5], making sure they were roughly at the same age and non-fake looking. In the second round, 7 people (4 females and 3 males; all at their mid-20s) were invited to rate the 50 faces on 2 dimensions: whether they fitted gender typicality (feminine for female faces and masculine for male faces), and whether they were good-looking.

After the second round, the 8 faces that received the most votes on each end of the dimensions were selected as the final set for the face-selection stage in the experiment. The 8 faces are demonstrated in Figure 2. Since these faces are all generated by AI from the same set of training data, they appear to be relatively homogeneous, however since the votes of these 7 people showed considerable consistency in both dimensions, the difference between these faces was detectable and could be noticed by the participants. All faces that were finally used as stimuli were evaluated as non-fake and natural.

2.2. Participants

After the experiment was published online, 262 participants completed the experiment. Among them, 148 were females, and 114 were males. Their ages range from 13 to 57 years old, and most are between 19 to 25 years old. 204 of the participants speak varieties of Mandarin at home, while other 58 speak some other languages such as Hmong or Wu Chinese.

As most participants fell between 19 to 25 years old and age was not the primary variable this study aims to investigate, participants younger than 17 and older than 29 years old were excluded to keep the distribution less complicated, giving us 249 final participants with 143 being females and 106 males.

3. RESULTS

The data obtained from the experiment include participant backgrounds, ratings of the cake, and faces they have selected.

3.1. Data Analysis

Participant backgrounds and the faces they have selected are coded into binary form to perform binomial logistic regression. The rating results are not transformed and are analyzed later separately.

Binomial logistic regression is performed to analyze whether the face's gender typicality can be predicted by the participant's gender, language background (Mandarin or not), speaker's gender, and the voice quality used in the passage. All independent variables are transformed into dummy variables. Voice quality is transformed into 2 dummy variables as there are 3 levels within it (breathy, modal, and creaky), and we would like to investigate each level's impact on stances. Modal voice is set to be the default and thus represented by "intercept" in the regression. The results are shown in Table 3.

	Coeffi-cient	SE	z-value	p-value
Intercept	0.75	0.44	1.71	0.0875
PG	-0.11	0.31	-0.35	0.7258
Language	0.33	0.36	0.92	0.3584
SG	1.39	0.33	4.23	**<0.0001**
Creaky	-1.11	0.38	-2.89	**0.0039**
Breathy	-0.53	0.37	-1.44	0.1496

Table 3: Results of binomial logistic regression against face's gender typicality. "PG" represents "participant's gender", and "SG" stands for "speaker's gender". P-values < 0.05 is emphasized in bold. Interactions are non-significant and omitted.

Another binomial logistic regression is performed to see if the good look of the selected face could be predicted by the participant's gender, language background (Mandarin or not), speaker's gender, and the voice quality used in the passage. The results are shown in Table 4.

	Coeffi-cient	SE	z-value	p-value
Intercept	1.25	0.42	2.97	**0.0030**
PG	-0.09	0.28	-0.31	0.7536
Language	0.04	0.33	0.13	0.8937
SG	-0.52	0.28	-1.85	0.0640
Creaky	-0.82	0.34	-2.42	**0.0157**
Breathy	0.05	0.34	0.14	0.8862

Table 4: Results of binomial logistic regression against face's good looks. "PG" represents "participant's gender", and "SG" stands for "speaker's gender". P-values < 0.05 is emphasized in bold. Interactions are non-significant and omitted.

The responses to the 3 statements about the cake are analyzed into frequencies of each rating and displayed in Figure 3.

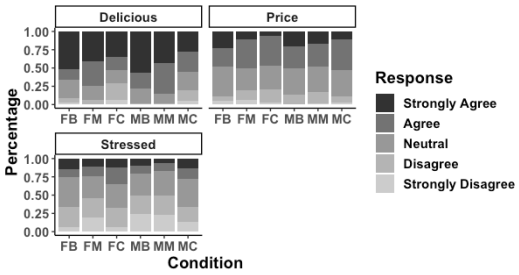

Figure 3: The percentage distribution of participants' responses to 3 statements about the cake after listening to the passage narrated by various gender and voice combinations. For example, "FB" represents "female breathy voice" and "MC" represents "male creaky voice".

As shown in Table 3, participant gender and language background are not related to the face's gender typicality. Keeping other factors the same, the odds of creaky voice being considered as fitting gender typicality is exp(-1.11) = 0.33 times of the odds of a non-creaky voice being considered as fitting gender typicality, which means that creaky voice is recognized as more gender-neutral to both genders. The regression also shows that the speaker's gender affects the gender typicality that participants select. This could be resulted from the pitch of audio stimuli; the female speaker had a deeper voice compared to average females, which could lead her voice being perceived as less feminine ([10]).

As to good looks, it is found that modal and creaky voices have significant impacts on the listener's judgment. While other variables are consistent, participants are exp(1.25) = 3.49 times more likely to select a good-looking face if they heard a modal voice than a non-modal voice. If they heard a creaky voice, on the other hand, the odds of selecting a good-looking face is exp(-0.82) = 0.44 times of hearing a non-creaky voice. It means that modal voice is perceived as more good-looking compared to non-modal voices, and creaky voice is less good-looking compared to modal and breathy voices.

Regarding the ratings of the statements about the cake, the most obvious trend shown in Figure 3 is that breathy voice gains the most positive judgments in deliciousness and price aspects while creaky voice gains the least. The voice quality does not affect how people perceive the cake as too much from such a friend and whether they feel stressed by the gift.

4. DISCUSSION

The fact that participants selected faces that fit gender typicality after hearing modal and breathy voices but not creaky voice is different from our expectation that, at least for male, creakiness indicate a certain degree of masculinity as it often co-occurs with low pitch, which naturally related to larger body size as claimed in [11], as well as previous studies in English, e.g. [12]. The difference between our results and anticipation agrees with the previous theory that gender is a performed sociocultural identity and is not fixed to biological sex [13]. When we view gender in terms of masculinity and femininity but not male and female, we are expressing the sociocultural ideologies about gender, which leads the genderization of linguistic forms [14]. In the current study, voice quality has been shown to be such a gendered linguistic form in Mandarin Chinese.

While gender typicality could be a rather neutral dimension, different stances perceived from these linguistic forms, or voice qualities, are much more subjective as they are reflected by ratings toward the cake and selection on the looks of the speakers. As to the stance toward the cake described in different voice qualities, the results have shown that the more spread the glottis is (which produces breathier voice), the higher the participants would rate the cake in terms of deliciousness and price. Deliciousness and expensiveness are two dimensions that reveal the stances that participants take toward the cake: breathier voice indirectly indexes more positive values in taste and price, and creaky voice indirectly indexes less positive values in taste and price. The looks of speakers, however, are thought to be the best while the voice is modal.

The limitations of this study are that the visual stimuli varied in only two dimensions in which the differences are not extremely obvious, and the experiment is only on the perception side so that it cannot allow us to observe the stances performed and exchanged between speakers like dialogues can do. Future studies may take on the results to design a production study and investigate the stance expressed by voice quality with a greater depth.

5. CONCLUSION

In this perception study, we have investigated the role of voice quality in perceiving stance in Mandarin. It is found that creaky voice will lead the listeners to construct an image of the speaker that is more gender-neutral and less good-looking, while modal voice is associated to more good-looking portrait. It is also found that the evaluation of an object can be affected by the voice quality of the narrator such that the more spread the glottis is, the higher the participants would rate the object in terms of subjective and objective values.

6. REFERENCES

[1] Habasque, Pierre. (2021). Is Creaky Voice a Valley Girl Feature? Stancetaking & Evolution of a Linguistic Stereotype. *Anglophonia.* 32. 10.4000/anglophonia.4104.

[2] Starr, R. L. (2015). Sweet voice: The role of voice quality in a Japanese feminine style. *Language in Society*, *44*(1), 1-34.

[3] Callier, P. R. (2013). *Linguistic context and the social meaning of voice quality variation*. Georgetown University.

[4] Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, *41*(1), 87-100.

[5] Du Bois, John W. (2007). "The stance triangle". In Engebretson, Robert (ed.). *Stancetaking in Discourse: Subjectivity, evaluation, interaction*. Amsterdam: John Benjamins.pp. 139-182. doi:10.1075/pbns.164.07du. ISBN 9789027254085.

[6] Bucholtz, M., & Hall, K. (2005). Identity and interaction: a sociocultural linguistic approach. Discourse Studies, 7(4–5), 585–614. https://doi.org/10.1177/1461445605054407.

[7]Corretge, R. (2022). *Praat Vocal Toolkit.* https://www.praatvocaltoolkit.com.

[8] Wright, R., Mansfield, C., & Panfili, L. (2019). Voice quality types and uses in North American English. *Anglophonia. French Journal of English Linguistics*, (27).

[9] a312863063 (2022). *New version of face generators based on StyleGAN2* [electronic resource: python source code]. github.com/a312863063/generators-with-stylegan2.

[10] Krahé, B., Uhlmann, A., & Herzberg, M. (2021). The voice gives it away: Male and female pitch as a cue for gender stereotyping. *Social Psychology, 52*(2), 101–113. https://doi.org/10.1027/1864-9335/a000441

[11] Pisanski, K., Isenstein, S. G., Montano, K. J., O'Connor, J. J., & Feinberg, D. R. (2017). Low is large: spatial location and pitch interact in voice-based body size estimation. *Attention, Perception, & Psychophysics*, *79*(4), 1239-1251.

[12] Lee, K. E. (2016). *The Perception of Creaky Voice: Does Speaker Gender Affect our Judgments?* The University of Kentucky.

[13] Butler, J. (2002). *Gender trouble*. routledge.

[14]Kiesling, S. F. (2019). The 'Gay Voice'and 'Brospeak': Towards a systematic model of stance.