Voice quality as a pitch-range indicator

Jianjing Kuang [1], *Yixuan Guo*[2], *Mark Liberman*[1]

[1] Department of Linguistics, University of Pennsylvania, U.S.A.
[2]AMCS, University of Pennsylvania, U.S.A.

kuangj@sas.upenn.edu, gyix@sas.upenn.edu, markyliberman@gmail.com

Abstract

Pitch perception plays a central role in processing speech prosody. Since f0 varies from speaker to speaker and from context to context, effective pitch-range normalization is thus important to uncover intended linguistic pitch targets. It has also been speculated that voice quality may play a role in pitch-range perception. Our previous study demonstrated that spectral balance indeed effectively affected the perception of pitch height: "tense voice", implemented as stimuli with spectral balance tilted towards higher frequency, was perceived as higher in pitch. Our previous study used non-speech stimuli, raising the possibility that listeners might not be in the speech mode; this current study therefore replicates the previous experiment using speech stimuli resynthesized with the same range of f0 contours and a similar spectral manipulation, and the same forced-choice pitch classification experiment with four spectral conditions. The results are consistent with our previous experiment: the pitch classification function was significantly shifted by differences in spectral balance. Listeners generally hear higher pitches when the spectrum includes more high-frequency energy (i.e., tenser phonation). Moreover, there is a salient perceptual bias: When the second peak is tenser, the effect is stronger. These new results further support the hypothesis that voice quality cues are strong indicators of pitch-range.

Index Terms: pitch perception, voice quality, f0, spectral slope

1. Introduction

Pitch perception plays a central role in processing speech prosody. Studies of pitch perception have primarily focused on f0 cues, since fundamental frequency (f0) appears to be the only acoustic correlate of pitch. However, since f0 range varies from speaker to speaker and from context to context, phonetic categories (e.g. tonal categories) thus overlap in acoustic signals. So effective pitch-range normalization is important to uncover intended linguistic pitch targets.

Studies [1-3] on speaker normalization have shown that listeners are able to identify the pitch location of very brief voice samples in an unknown speaker's range in the absence of any contextual cues. This suggests that listeners must use other signal-internal information that co-varies with f0 as cues to pitch range.

Both [1] and [2] speculated that voice quality could be such a cue. Indeed, co-variation between f0 and voice quality has been found in pitch production studies (singing [4-7]; speech [8]): The lowest pitch range is associated with vocal fry, and the highest pitch range is associated with tense voice and falsetto. The question is then whether this co-variation also occurs in the perception domain.

Our previous study [9] thus tested the hypothesis whether the presence of tense voice can facilitate the perception of high pitch. It has been well established (see [10] for a review; cross-linguistic studies [11-22]) that spectral slope of the voice source spectrum is an important indicator of voice quality: a relatively steep spectral slope is associated with a breathier voice and that a flat spectral slope is associated with a tenser or creakier voice (note that the latter also includes pulse-to-pulse variability). Therefore, "Tense voice" was implemented as stimuli with spectral slope tiled towards higher frequency. Four sets of synthetic overtone series with different spectral conditions were used in a pitch classification experiment. We found that indeed listeners significantly perceived more high pitch in the "tense voice" condition. This was true for both tonal and non-tonal speakers [23], so integrating spectral slope into pitch perception appears to be a universal psychoacoustic mechanism.

However, since the previous experiment used non-speech stimuli, some doubt was raised: listeners might not be in the speech mode in the previous experiment. And it is possible that listeners can ignore the spectral cues in a speech task. Studies have shown that listeners behaviors differently in processing speech and non-speech stimuli (e.g., [24,25]). Neural imaging study also demonstrated that people use different part of brain to process linguistic pitch and non-linguistic pitch (e.g., [26]). Therefore, it is important to validate our result with speech stimuli.

2. Method

2.1. Stimuli

Same as the previous experiment, the goal was to create four sets of utterances with two f0 peaks, and the two f0 peaks vary in spectral conditions. In this current experiment, each peak was carried by three /ma/ syllables, so that the whole sequence had the prosodic pattern of a phrase like "phonetic condition" or "electric banana". The stimuli were resynthesized from the natural production of a male English speaker. The speaker was asked to produce tokens of /ma.ˈma.ma/ with the same intonation pattern as "two twenty".

In order to preserve the naturalness of the original utterance, we chose to use TANDEM-STRAIGHT algorithm [27] for resynthesis. In the TANDEM-STRAIGHT algorithm, a single token of /ma.'ma.ma/ was first analyzed into three components: the f0 contour (f0(t)), the STRAIGHT spectrogram (S(f,t)), and the aperiodicity component (A(f,t)). Then the three components were modified with 22 sets of parameters (11 f0 steps multiplied by 2 spectral tilt values).

For f0 manipulation, a Hann function (a cosine period with the peak in the middle) was used for each peak. The base was the same (at 120 Hz) for both peaks; and the maximum value of the first peak (169.34 Hz) was kept the constant for all stimuli, while the second peak was an 11-step continuum varying from 153.06 Hz to 187.36 Hz (0.35 semitone/step). The result is shown in Figure 1. This setting is the same as our previous study [9].

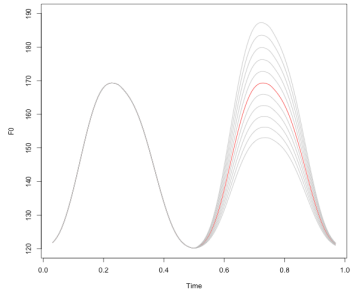
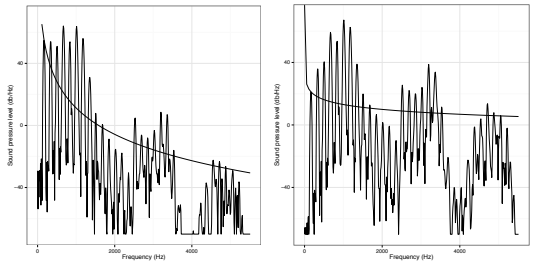


Figure 1*: f0 manipulation: the first peak has a constant f0 value at 169.34 Hz, and the second peak is a continuum with 11 steps. Peaks 1 and 2 are identical at step 6 (red/dark lines for the second peak).*

To manipulate voice quality cues, two versions of spectral balance were created: one with relatively more high-frequency energy (i.e. tenser version) and one with relatively less high-frequency energy (i.e. breathier version). The breathier version was the original spectrum of the natural production, while the tenser version was modified so that the Fourier spectrum was 6 dB/octave greater than the breathier version. This modification corresponded to a differentiation operation of the Fourier spectrum, due to the derivative-differentiation property of Fourier transform, therefore, it is precisely equivalent to the spectral contrast in our previous study. The result of this spectral boost is depicted in Figure 2. In a second experiment, we only boosted the high frequency component of the tense version by 3 dB/octave.

.



(a) (b)

Figure 2*: Spectral manipulation: (a) original; (b) boosted*

Finally, the modified parameters were combined and resynthesized into 22 tokens of different f0 peaks (11 steps) and spectral slope (2 values). These single peaks were 0.52 seconds in duration and then concatenated to create 4 sets of two-peak stimuli, labeled with letter A-D in the same way as previous (44 stimuli in total). Thus there were 4 different spectral conditions in the stimuli (implied phonation types are presented in brackets in relative terms):

- Set A: Both peaks have the original spectrum (i.e., breathier + breathier)
- Set B: Both peaks have the boosted spectrum (i.e., tenser + tenser)
- Set C: The first peak has the original spectrum, and the second has the boosted spectrum, with a 200 ms transition in the middle (i.e., breathier + tenser)
- Set D: The first part has the boosted spectrum, and the second has the original spectrum, with a 200 ms transition in the middle (i.e., tenser + breathier).

Therefore, there were 44 stimuli (11 f0 steps x 4 spectral conditions) in a total. All stimuli were 1.05 s in duration.

2.2. Procedure

A forced-choice pitch classification task was used to test listeners' categorization of pitch values under different spectral conditions. Five copies of each stimulus were presented in random order to each listener. For each trial, the listeners were asked to focus on pitch and to evaluate whether the second "maMAma" word was higher or lower than the first one by clicking on the corresponding buttons on the computer screen. To introduce the idea of pitch to an English speaker, we used the examples of English intonation. For example, the phrase "my name" is higher in "Anna may know my name?" than in "Anna may know my name." In the practice session, examples from set A were used to demonstrate the task. This was to make sure that listeners would attend to pitch difference but not other cues (e.g. intensity). The experiment was run with Qualtrics online survey system. The subjects were instructed to use headphones or earbuds to do the experiment.

2.3. Subjects

English speakers between age 18 and 22 were recruited from the student population at the University of Pennsylvania. There are 34 listeners in the first experiment, and 30 listeners in the second experiment. There is no overlapping between the two subject pools. All the subjects reported to have normal hearing and speaking.

3. Results

3.1. Experiment 1

Figure 3 shows the proportion of "peak 2 is higher" responses for all English listeners. The main effects of spectral conditions were evaluated using an MCMC generalized linear mixed-effects model (*mcmcglmm* package in R). f0 steps (1-11) and spectral conditions (A, B, C and D) were used as fixed factors, and random intercepts and slopes were included as subjects. The main effects of the spectral conditions are summarized in Table 1. The results are reported as means of regression coefficients followed by 95% highest posterior density intervals in square brackets and associated p-values. As shown in Table 1, the results demonstrate that pitch classification functions significantly shifted in set C and D.

Overall, the current experiment successfully replicates the result from the previous experiment (shown in Figure 4): the perception of pitch height was strongly biased by spectral

cues. As shown in Figure 3, compared with set A and B, where the two peaks have identical spectral conditions, the pitch classification function for set C (breathier + tenser combination) was dominated by "peak 2 is higher" responses; by contrast, the pitch classification function of set D (tenser + breathier combination) shifted in the opposite direction. In other words, when the second peak was tenser than the first, the second peak tended to be perceived as a higher pitch, and when the second peak was breathier than the first, the second tended to be perceived as a lower pitch.

	A	B	C
B	0.17[-0.05,0.44] p=0.17		
C	1.1[0.5,1.9] p<0.001	1.03[0.7,1.4] p<0.001	
D	0.5[0.3,0.7] p<0.001	0.4[-0.7,-0.1] p<0.001	1.5[-2.1,-0.9] p<0.001

Table 1. *Main effects of spectral conditions for every pair of conditions. Means of regression coefficients followed by 95% highest posterior density intervals in square brackets and associated p-values.*

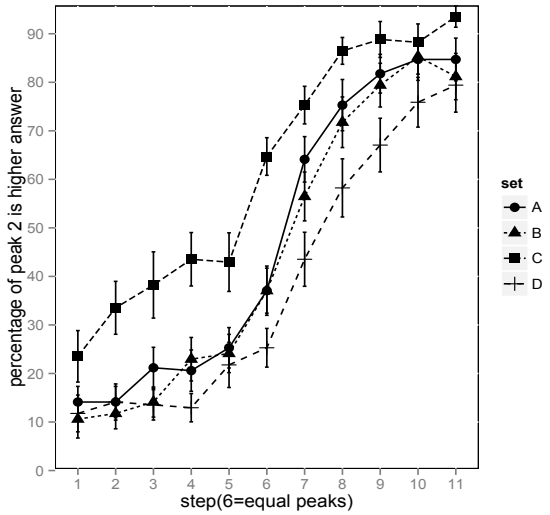


Figure 3*: Pitch classification functions for English listeners. X-axis=f0 steps, y-axis=proportion of "peak 2 is higher" responses; line patterns denote different spectral conditions. Error bars denote 95% confidence intervals.*

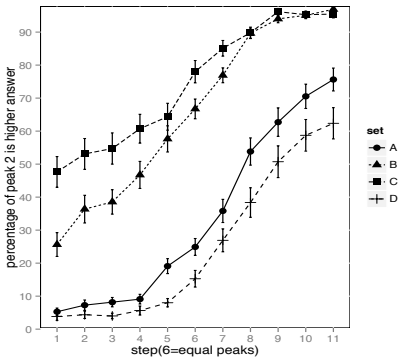


Figure 4*: From previous study: Pitch classification functions for English listeners with the non-speech stimuli.*

It is worth noting that there are some differences apparently due to the nature of the stimuli. In the non-speech stimuli version of the experiment (Figure 4), set B also significantly shifted from set A. This means that the spectral condition of the second peak itself has a strong effect. However, in the speech stimuli version (Figure 3), set B no longer shifts from set A, which suggests that listeners are insensitive to the absolute quality of the utterance, but cared more about the relative difference between the two peaks. Our previous study also found salient individual differences: some listeners only used spectral cues or f0 cues in the task. But we didn't find such variation in the current experiment, which suggests that integration of the two cues is more obligatory in a speech-processing task.

In addition, although both set C and D significantly shift, the effect of set C (breathier + tenser) is greater than set D (tenser + breathier). This suggests a perceptual bias: when the second peak is tenser, the effect is stronger.

Since listeners appeared to be very sensitive to spectral difference, we wondered if we can still replicate this effect when reducing the spectral difference between the two peaks. Therefore, we performed another experiment in which the difference between the two spectral conditions was only half as great as in the first experiment.

3.2. Experiment 2

The stimuli and procedure of experiment 2 are exactly the same, except that the spectral difference was only 3dB/octave, half of the 6 dB/octave use in experiment 1. Another 30 listeners were recruited from the student population to participate in the experiment.

Table 2 is the summary of the main effects of the spectral conditions. Similar to Table 1, Set C and set D significantly shift from set A and B. This effect can be clearly seen in Figure 5.

Although the spectral difference is much smaller in the second experiment, the salience of the effect remains, as shown in Figure 5. This means that listeners are very sensitive to the spectral difference. Moreover, Figure 3 and Figure 5 have very similar amount of shift, so it seems that greater spectral difference does not introduce more shift, at least in the range of values tested.

	A	B	C
B	0.01 [0.18,0.22] P=0.9		
C	0.7[0.5,0.9] p<0.001	1.7[0.9,2.5] p<0.001	
D	0.45[0.24,0.67] p>0.001	1[0.5,1.7] p<0.001	1.4[1.0, 1.8] p<0.001

Table 2. *Main effects of spectral conditions for every pair of conditions. Means of regression coefficients followed by 95% highest posterior density intervals in square brackets and associated p-values.*

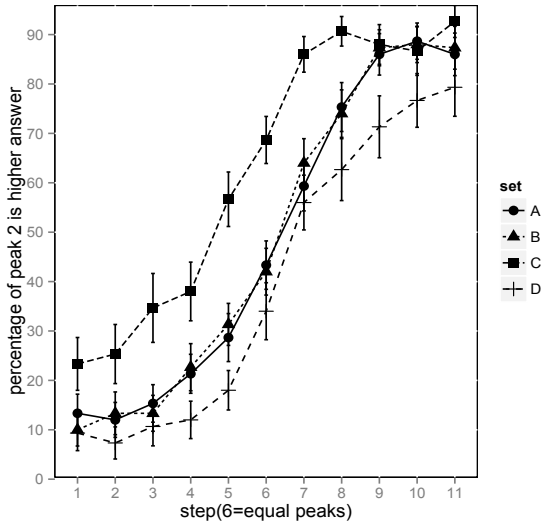


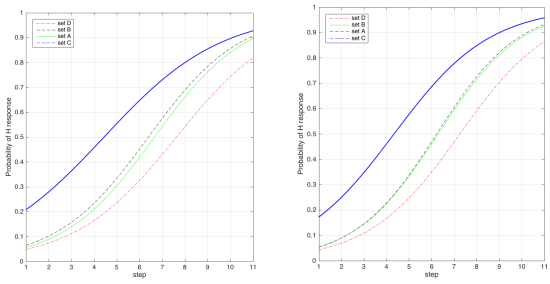
Figure 5*: Pitch classification functions for English listeners. Spectral difference is 3dB/octave. X-axis=f0 steps, y-axis=proportion of "peak 2 is higher" responses; line patterns denote different spectral conditions. Error bars denote 95% confidence intervals.*

To quantify the amount of shift, we fitted the classification functions with a sigmoid function to determine the threshold (alpha; i.e., left-to-right shift) and the slope (beta) of the response probability. Figure 6 exhibits the fitted curves of the responses from experiments 1 and 2, and table 3 shows the values of threshold (alpha) and slope (beta).

As we can see here, experiment 1 and 2 have very similar results, despite the fact that the spectral difference between two peaks in the experiment 2 is much smaller. In both experiments, set C (breathier + tenser) shifts two steps from A and B to the left (i.e., alpha.set C – alpha.set B/A = -2), indicating that by manipulating the spectral condition, the stimuli sounded 0.7 semitone (0.35 semitone/step x 2) higher to the listeners. On the other side, set D (tenser + breathier) shifts one step from A and B to the right (i.e., alpha.set D – alpha.set B/A=1), indicating that this set of stimuli sounded 0.35 semitone lower to the listeners.

Set	**Exp1**		**Exp2**	
	α	**β**	**α**	**β**
A	6.4	2.0	6.3	1.9
B	6.6	2.0	6.2	1.8
C	4.4	2.5	4.3	2.1
D	7.6	2.2	7.2	2.0

Table 3. *Threshold (*α*) and slope (β) of the fitted sigmoid function.*



(a) (b)

Figure 6: *Fitted curve for experiment 1 (a) and 2 (b). blue=set C; black= set B; green= set A; red= set D*

4. Discussion

This study used resynthesized speech stimuli to replicate our earlier finding that voice quality cues contribute strongly to the perception of relative (peak) pitch in stimuli with rising-falling f0 glides similar to those typical in speech. Listeners generally perceive a higher relative pitch for a peak where higher-frequency components in the spectrum have more energy (indicating a tenser voice quality [10]), when they are comparing it to a peak with the same fundamental frequency but less high-frequency energy. The direction of the shift is consistent with the co-varying relationship between f0 and voice quality: high f0 is naturally produced by a tense voice [6]. This study had two new findings: 1) We tested the robustness of the spectral cue by reducing the spectral difference between the two f0 peaks, and found that even a relatively small spectral difference can lead to significant shift of the pitch classification function, suggesting that listeners are very sensitive to the voice quality cue. 2) Speech mode does have an effect the behavior of the pitch perception. When listeners in the speech mode, they are more likely to integrate both f0 and spectral cues in pitch perception, and they are less sensitive to the absolute quality of the utterance. In sum, this study thus further supports the hypothesis that voice quality cues and f0 are integrated in pitch perceptions, perhaps because voice quality is a strong indicator of pitch range.

The findings of this study have important implications for prosody studies: pitch is not merely f0, either in production or in perception. As suggested in this study, pitch perception can be determined by both f0 and voice quality cues. Thus, what is perceptually "higher" does not necessarily have a higher f0 in the signal. The importance of voice quality in speech prosody has been received more and more attention, especially in the paralinguistic level (e.g., emotional speech) (e.g., [28]); voice quality is the enhancement cue for tonal contrasts [29]; it is sensitive to prosodic structures (e.g., [30]). But we show in this study that voice quality plays a very fundamental role in prosodic structure, as it is a part of pitch processing. Pitch analysis and synthesis thus should take voice quality cues into account. Previous study has shown that tone classification can be achieved with spectral information only [31].

This finding is instrumental to speaker normalization, as voice quality can provide information on pitch location for a speaker; for example, a tense voice indicates that the speaker has nearly reached his/her highest range (or is speaking at his/her highest pitch).

5. References

[1] D. Honorof and D. Whalen, "Perception of pitch location within a speaker's f0 range," *J. Acoust. Soc. Am.,* vol. 117, pp. 2193-2200, 2005.

[2] C.-Y. Lee, "Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study," *J. Acoust. Soc. Am.,* vol. 125, pp. 1125-1137, 2009.

[3] J. Bishop and P. Keating, "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex," *J. of the Acoust. Soc. Am.,* vol. 132, pp. 1100-1112, 2012.

[4] H. Hollien and J. F. Michel, "Vocal fry as a phonational register," *Journal of Speech and Hearing Research* vol. 11, p. 600 1968.

[5] H. Hollien, "On Vocal registers," *Journal of Phonetics* vol. 2, pp. 125-143 1974.

[6] I. R. Titze, "A framework for the study of vocal registers," *Journal of Voice* vol. 2, pp. 183-194 1988.

[7] B. Roubeau, N. Henrich, and M. Castellengo, "Laryngeal Vibratory Mechanisms: The Notion of Vocal Register Revisited," *Journal of Voice,* vol. 23, pp. 425-438, 2009.

[8] J. Kuang, The covariation between pitch and phonation: creaky voice in Mandarin tones. The 89th Annual Meeting of the Linguistic Society of America, 2015.

[9] J. Kuang and M. Liberman, "Influence of spectral cues on the perception of pitch height", Proceeding of ICPH 18, 2015.

[10] C. Gobl and A. Ní Chasaide, "Voice source variation," in *The Handbook of Phonetic Science*, W. J. Hardcastle and J. Laver, Eds., ed Oxford: Blackwell, 2012, pp. 378-423.

[11] J. E. Andruski, "Tone clarity in mixed pitch/phonation-type tones," *Journal of Phonetics,* vol. 34, pp. 388-404, 2006.

[12] J. E. Andruski and M. Ratliff, "Phonation types in production of phonological tone: the case of Green Mong," *Journal of the International Phonetic Association,* vol. 30, pp. 37-61, 2000.

[13] B. Blankenship, "The timing of nonmodal phonation in vowels," *Journal of Phonetics,* vol. 30, pp. 163-191, 2002.

[14] A. S. Abramson, T. Luangthongkum, and P. W. Nye, "Voice register in Suai (Kuai): An analysis of perceptual and acoustic data," *Phonetica,* vol. 61, pp. 147–171, 2004.

[15] E. Thurgood, "Phonation types in Javanese," *Oceanic Linguistics* vol. 43, pp. 277–295, 2004.

[16] A. L. Miller, "Guttural vowels and guttural co-articulation in Ju|'hoansi," *Journal of Phonetics,* vol. 35, pp. 56-84, 2007.

[17] C. T. DiCanio, "The phonetics of register in Takhian Thong Chong," *Journal of the International Phonetic Association,* vol. 39, pp. 162-188, 2009.

[18] C. M. Esposito, "Variation in contrastive phonation in Santa Ana Del Valle Zapotec," *Journal of the International Phonetic Association,* vol. 40, pp. 181-198, 2010.

[19] M. Garellek and P. Keating, "The acoustic consequences of phonation and tone interactions in Jalapa Mazatec," *Journal of the International Phonetic Association,* vol. 41, pp. 185-205, 2011.

[20] J. Kuang and P. Keating, "Glottal articulations in tense vs. lax phonation contrasts," *J. Acoust. Soc. Am.,* vol. 136, pp. 2784–2797, 2014.

[21] C. M. Esposito, "An acoustic and electroglottographic study of White Hmong phonation," *Journal of Phonetics,* vol. 40, pp. 466-476, 2012.

[22] S. D. Khan, "The phonetics of contrastive phonation in Gujarati," *Journal of Phonetics,* vol. 40, pp. 780-795, 2012.

[23] J. Kuang and M. Liberman, "Effect of spectral slope on pitch perception", Interspeech 2015.

[24] A. M. Liberman, "Some characteristics of perception in the speech mode." *Perception and its Disorders* vol. 48, pp. 238-254. 1970.

[25] B. H. Repp, "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," *Psychological Bulletin*, vol. *92*, pp. 81, 1982.

[26] J. Merrill, D. Sammler, M. Bangert, D. Goldhahn, G. Lohmann, R. Turner, and A. D. Friederici. "Perception of words and pitch patterns in song and speech." *Frontiers in psychology* 3, 2012.

[27] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation." Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference, pp. 3933-3936. IEEE, 2008

[28] C. Gobl and A. Ní Chasaide. "The role of voice quality in communicating emotion, mood and attitude." *Speech communication*, vol. 40, pp. 189-212, 2003.

[29] J. Kuang, "The tonal space of contrastive five level tones," *Phonetica,* vol. 70, pp. 1-23, 2013.

[30] M. Garellek. "Voice quality strengthening and glottalization." *Journal of Phonetics,* vol. 45, pp.106-113, 2014.

[31] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly Accurate Mandarin Tone Classification In The Absence of Pitch Information," in *International conference on Speech Prosody*, Dublin, pp. 673-677, 2014.

.