interspeech 2008 incorporating SST 08

Acoustic Analysis of Imitated Voice Produced by a Professional Impersonator

Tatsuya Kitamura

Department of Intelligence and Informatics,
Faculty of Intelligence and Informatics, Konan University, Kobe, Japan
t-kitamu@konan-u.ac.jp

Abstract

We conducted a comparative study of a voice produced by a professional impersonator imitating a target speaker to explore the acoustical characteristics that the impersonator changes from his natural voice to imitate a target voice. Comparison of the pitch frequency of the target and imitated voices showed that the mean and dynamics of the pitch frequency of the imitated voice are changed so that they become closer to those of the target voice. The spectra of vowels uttered by the speakers are also similar in their shape and formant frequencies. In the imitated voice, the formant frequencies shift by up to 68 % from those of the impersonator's natural voice. Moreover, the differences between the amplitudes of the first and second harmonics (H1-H2), a measure of the glottal source characteristics, of the target and imitated voices are small. These results clearly demonstrate that the impersonator controls vocal tract acoustic characteristics as well as those of the glottal source and pitch frequency to imitate the target voice.

Index Terms: imitated voice, impersonator, pitch frequency, vocal tract acoustic characteristics, glottal source characteristics

1. Introduction

Impersonators, who can closely imitate other speakers' voices, are an intriguing subject of speech science. Which acoustic characteristics do they adjust to those of the target voice? How do they control their speech production system while imitating a voice? How can we find that the imitated voice resembles the target one? The answers to these questions may offer fruitful insights into studies on the perception of speaker individuality and voice conversion in synthetic speech. Except for a few previous studies [1]-[4], however, imitated voices have not been studied, probably because of the difficulty of obtaining subjects who can dexterously imitate voices. The aim of the present study is thus to explore the acoustical characteristics that a professional impersonator changes from his natural voice to imitate a target voice.

In an earlier study, Suzuki [1] analyzed an imitated utterance by a professional impersonator and showed that some properties appeared in spectrograms that are clearly different from the target utterance, although a human ear cannot perceive these differences. Recently, Zetterholm [2]-[4] has carried out auditory and acoustic analyses of imitated utterances. She demonstrated that a professional impersonator captures the speech style, dialect, pronunciation, and intonation pattern. These studies have revealed the acoustic characteristics and speech behavior captured by impersonators to imitate target voices; however, we have not yet answered the above questions. In the present study, we therefore make another attempt to analyze an imitated voice.

In this paper, we describe the results of the acoustical analyses of a target speaker's voice, the voice imitated by a professional impersonator, and the impersonator's natural voice, and speculate on how he controls his speech production system to imitate the target voice.

2. Speech data

A target speaker (a professional comic storyteller) and a professional impersonator recorded the following two sentences at a sampling rate of 16 kHz with 16-bit resolution. In this study, the former and latter speakers are referred to as Speakers A and B, respectively. Speaker B recorded the sentences in the imitated voice and his natural voice on separate days.

Sentence 1 Ichidode i:kara mitemitai, nyo:boga hesokuri kakusutoko. (Just once, I wish to catch my wife hiding her secret savings.)

Sentence 2 Dekakeru nekoni yukisaki kikeba, ryoko:ga sukide mata tabida. (Asked where to go, the cat replied that he is going on a travel again because he loves to. (This Japanese sentence has a humorous play on words.))

When Speaker B imitated Speaker A's voice, he spoke immediately after Speaker A's utterances, that is, he followed the target utterances. Speaker B recorded his natural voice alone on another day. Each sentence was produced twice (Speaker B produced each sentence twice in the imitated voice and his natural voice.). We used one sample per sentence in the analysis, and thus analyzed three utterances for each sentence.

3. Analysis method

To explore the acoustical characteristics that Speaker B controlled to imitate Speaker A, we measured (1) the pitch frequency, (2) discrete Fourier transform (DFT) spectra, (3) formant frequencies, (4) the difference between the amplitudes of the first and second harmonics (H1-H2) [5], and (5) the syllable duration, and compared them between the speakers and between the manner of speech (imitated and natural voicing). The power of the speech waves was excluded from the analysis since the speech data was affected by background noise and the reverberation of the room.

The pitch frequency was extracted using the "Pitch Contour" function of WaveSurfer [6] with its default parameters; the method used was ESPS [7], the frame length was 7.5 ms, and the frame period was 10 ms. The obtained pitch frequency was corrected manually.

The DFT spectra were calculated from overlapping Hanning window frames; the frame length was 64 ms with an 8 ms frame period. The first, second, third, and fourth formant frequencies were extracted from spectral envelopes obtained

from 40-order DFT cepstra by a peak-picking method. H1-H2, a measure of the glottal source characteristics, was measured from the DFT spectra.

Syllable duration was calculated from the syllable boundaries delimited manually referring to the spectrograms of the speech data. The correlation coefficient r was then calculated for the syllable duration in the utterances.

4. Results

4.1. Pitch frequency

The mean pitch frequency of the speech data of Sentence 1 uttered by Speaker A is 167.2 Hz, that uttered by Speaker B in the imitated voice is 185.1 Hz, and that uttered by Speaker B in his natural voice is 152.0 Hz. The mean pitch frequency of the speech data of Sentence 2 uttered by Speaker A is 162.2 Hz, that uttered by Speaker B in the imitated voice is 187.3 Hz, and that uttered by Speaker B in his natural voice is 154.8 Hz. The mean pitch frequencies do not exhibit much difference between the two sentences for each speaker. The mean pitch frequency of each imitated utterance is approximately 20 Hz higher than that of each target utterance implying that Speaker B exaggerates Speaker A's high-pitched voice.

Figure 1 shows the pitch frequency contours of the speech data in a logarithmic scale. The arrows and numbers in the figures represent the tilt of the pitch frequency contours. The positive values represent a rising pitch frequency and the negative values represent a falling pitch frequency. The shape of the pitch frequency contours and the tilt, particularly in the first half of Sentence 1 (Fig. 1(a)), of the imitated utterances are more similar to those of the target utterances than those of the Speaker B's natural utterances.

4.2. DFT spectrum

The DFT spectra of a long vowel /i:/ and vowel /a/ in the first half of Sentence 1 are shown in Fig. 2. The spectra are averaged over the frames. The shape of the DFT spectra of Speaker B's imitated vowels (the middle panels of the figures) closely resembles that of Speaker A's vowels (the top panels of the figures) not only in the spectral peaks but also in the spectral dip in the frequency region from 3.0 to 3.4 kHz.

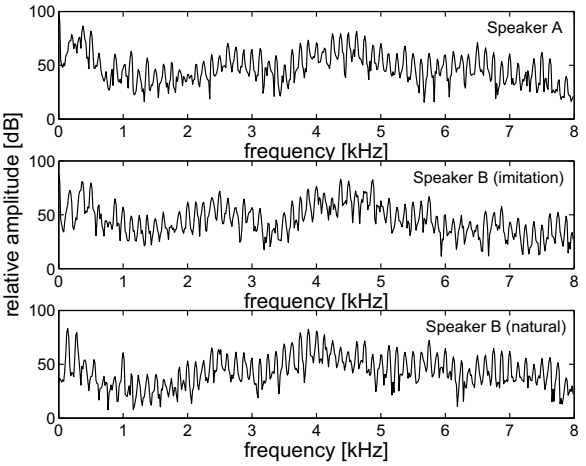
On the other hand, the shape of the DFT spectra of Speaker B's imitated vowels differs significantly from that of his natural vowels. These results clearly demonstrate that the impersonator changes the shape (and probably also the length) of his vocal tract to adjust the frequency and bandwidth of the spectral peaks and dip while imitating a voice.

4.3. Formant frequencies

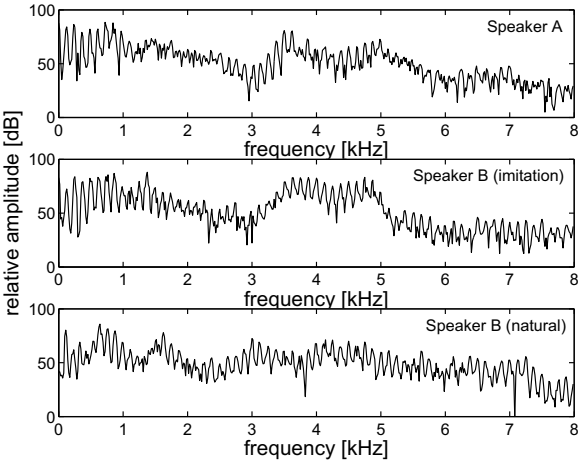
The first (F1), second (F2), third (F3), and fourth (F4) frequencies of the long vowel /i:/ and vowel /a/ are shown in Tables 1 and 2, respectively. They are measured from the envelopes of the DFT spectra (Figs. 2(a) and (b)).

To examine the similarity of the formant frequencies of Speaker A's vowels and Speaker B's imitated vowels, we calculated the percentage differences between each of the formant frequencies. For the long vowel /i:/, the difference for F1 is 4 %, that for F2 is 8 %, and that for F4 is 0 %. For the vowel /a/, the difference for F1 is 4 %, that for F2 is 9 %, that for F3 is 0 %, and that for F4 is 1 %. These results show that the differences between the formant frequencies are within 4 % except for F2.

Likewise, we analyzed the differences between the formant



(a) long vowel /i:/



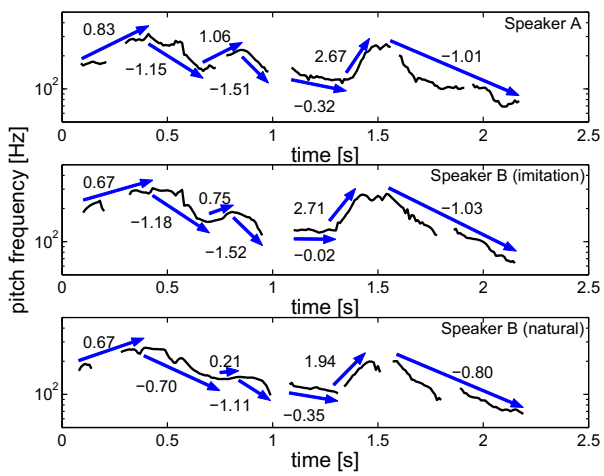
(b) vowel /a/

Figure 2: DFT spectra of a long vowel /i:/ and vowel /a/ in the first half of Sentence 1. The top panels show the spectra of Speaker A's vowels, the middle ones show those of Speaker B's imitated vowels, and the bottom ones show those of Speaker B's natural vowels.

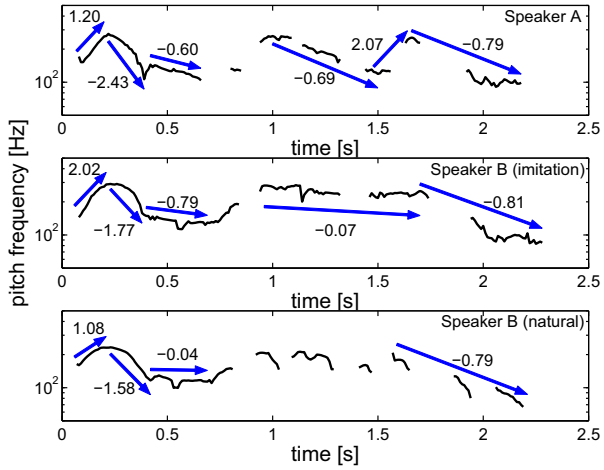
frequencies for Speaker B's natural and imitated vowels. For the long vowel /i:/, the difference for F1 is 62 %, that for F2 is 1 %, and that for F4 is 0 %. For the vowel /a/, the difference for F1 is 19 %, that for F2 is 12 %, that for F3 is 17 %, and that for F4 is 4 %. These results show that the differences between the formant frequencies are as high as 62 % and that Speaker B's natural F4 is close to that of Speaker A.

4.4. H1-H2

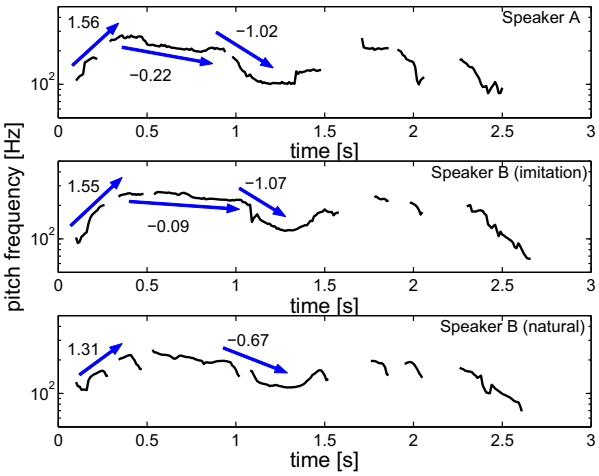
We measured H1-H2 from the DFT spectra of the long vowel /i:/ (Fig. 2(a)) and vowel /a/ (Fig. 2(b)). For the former vowel, H1-H2 for the target voice is -8.03 dB, that for the imitated voice is -8.12 dB, and that for Speaker B's natural voice is 3.82 dB, showing that the values of the target and imitated voices are exceedingly close. For the latter vowel, H1-H2 for the target voice is -2.22 dB, that for the imitated voice is -11.55 dB, and that for Speaker B's natural voice is 18.3 dB. In this case, the signs of



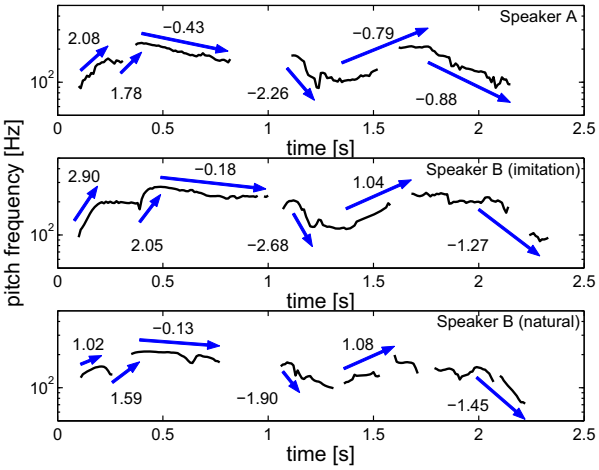
(a) the first half of Sentence 1



(b) the second half of Sentence 1



(c) the first half of Sentence 2



(d) the second half of Sentence 2

Figure 1: Pitch frequency contours of the first and second halves of Sentences 1 and 2 with arrows and numbers representing their tilt. The left panels (a and b) show the results for Sentence 1 and the right panels (c and d) show those for Sentence 2. The top panels show the pitch frequency contours of Speaker A's utterances, the middle ones show those of Speaker B's imitated utterances, and the bottom ones show those of Speaker B's natural utterances.

Table 1: The first, second, third, and fourth formant frequencies in Hz of a long vowel /i:/ in the first half of Sentence 1.

	F1	F2	F3	F4
Speaker A	390.1	2687.5	3171.9	4031.2
Speaker B (imitation)	406.2	2468.8	—	4049.9
Speaker B (natural)	250.0	2484.4	2906.2	4031.2

Table 2: The first, second, third, and fourth formant frequencies in Hz of a vowel /a/ in the first half of Sentence 1.

	F1	F2	F3	F4
Speaker A	765.6	1484.4	3578.1	4000.0
Speaker B (imitation)	796.9	1343.8	3593.8	3968.8
Speaker B (natural)	671.2	1531.2	3078.1	4140.6

the values for the target and imitated voices are identical (negative) while the value for Speaker B's natural voice is positive. These results indicate that the impersonator adjusts the glottal source characteristics to those of the target speaker during voice imitation.

4.5. Syllable duration

Figure 3 depicts the syllable duration of the first and second halves of Sentence 1. The syllable duration shows better agreement between Speaker A's voice and Speaker B's natural voice than between Speaker A's voice and Speaker B's imitated voice. The correlation coefficient r between the syllable duration of Speaker A's voice and Speaker B's natural voice is 0.847 for

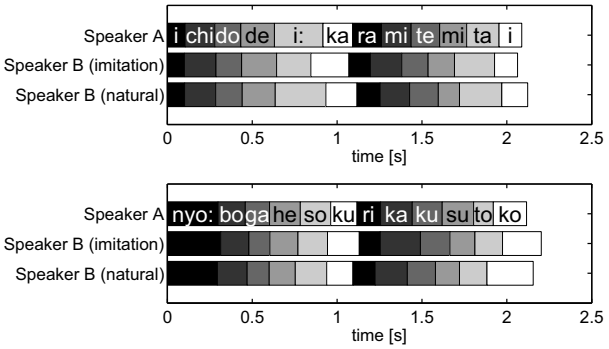


Figure 3: Syllable duration of the first (upper panel) and second (lower panel) halves of Sentence 1.

Sentence 1 and 0.758 for Sentence 2, and the correlation coefficient between the syllable duration of Speaker B's natural and imitated voices is 0.755 for Sentence 1 and 0.629 for Sentence 2. The results indicate that the impersonator does not adjust his syllable duration to that of the target speaker; it is, however, undeniable that the impersonator exaggerates the syllable duration of the target speaker. In addition, the results do not mean that the syllable duration cannot contribute to the perception of speaker individuality.

5. Discussion

The acoustic character of a speaker can be assumed to be coded in the brain as a vector in a multidimensional space of acoustic characteristics with its origin at the average of the characteristics. The origin of the space and the relative perceptual contribution of each acoustic characteristic may have a certain level of commonality, if they are not identical, among individuals. Impersonators probably imitate a target voice by adjusting the vector of their voice to that of the target voice. In the present study, the impersonator changes the mean and dynamics of his pitch frequency, the shape of his speech spectra, his formant frequencies (F1, F3, and F4), and his glottal source characteristics, which may show greater differences from the origin.

The mean pitch frequencies of the imitated utterances are approximately 20 Hz higher than those of the target voice in the two sentences. Speaker A, the target speaker, has a distinctive high-pitch frequency for a male. Thus, Speaker B, the professional impersonator, probably exaggerates this acoustic characteristic. Similar results have been reported by Zetterholm [2].

Speaker B imitates the shape of the pitch frequency contour in many parts of his utterances. Akagi and Ienaga [8] demonstrated that the dynamics of the pitch frequency contour as well as the mean pitch frequency contributes the perceptual speaker identification of three-mora words in psychoacoustic experiments. They reported that the perceptual contribution of the former is greater than that of the latter. According to their results, it is reasonable that the impersonator imitates the dynamics of the pitch frequency as well as the mean.

Speaker B also imitates the shape of the speech spectra including F1, F3, and F4 but excluding F2. This result may imply that F2 only has a small perceptual contribution to speaker identification or that the impersonator has difficulty adjusting F2 to the target value. Both possibilities need to be examined.

In the DFT spectra of the imitated vowels, there is a dip in the frequency region from 3.0 to 3.4 kHz. Spectral dips in general have no auditory effect but may contribute to the adjustment of the spectral shape to the target shape. A potential source of the dip is the piriform fossae, a pair of bilateral cavities located behind the laryngeal tube. The sinuses produce one or two dips in speech spectra [9] and their deformation affects the frequency and bandwidth of the dips [10].

In addition to the acoustic characteristics mentioned above, the impersonator changes his glottal source characteristics (H1-H2) during imitation. The target speaker, aged 70, has a hoarse voice characteristic of elderly persons. The result for H1-H2 possibly shows that the impersonator tried to imitate the hoarse voice.

6. Conclusions

In this study, we compared the acoustic characteristics between a target voice, an imitation of the target voice, and the impersonator's natural voice. The results revealed that the importance of the mean and dynamics of the pitch frequency, the vocal tract acoustic characteristics, and the glottal source characteristics when imitating a voice, although the results depend on the combination of the target speaker and impersonator.

Our future work is to examine the effects of acoustic characteristics on the perception of speaker individuality and measure the morphological differences of the speech production system between imitated and natural phonation.

7. Acknowledgements

This research was partly supported by Strategic Information and Communications R&D Programme (SCOPE) (No. 071705001) of the Ministry of Internal Affairs and Communications, Japan.

8. References

[1] Suzuki, M., "Spectrograms: speaker identification from the viewpoint of voice imitation," Gengo-seikatsu, 207, 37–41 (1968).

[2] Zetterholm, E., "Impersonation: reproduction of speech," Working Papers, Dept. of Linguistics, Lund University, 49, 176–179 (2001).

[3] Zetterholm, E., "A comparative survey of phonetic features of two impersonators," TMH-QPSR, 44, 129–132 (2002).

[4] Zetterholm, E., "Same speaker: different voices: A study of one impersonator and some of his different imitations," Proc. Int. Conf. Speech Sci. & Tech., 70–75 (2006).

[5] Hanson, H.M., Stevens, K.N., Kuo, H.-K. J., Chen, M.Y., and Slifka, J.,"Towards models of phonation," J. Phonet., 29, 451-480 (2001).

[6] WaveSurfer, `http://www.speech.kth.se/wavesurfer/`

[7] Talkin, D.,"A robust algorithm for pitch tracking (RAPT),"In W.B. Kleijn and K.K. Paliwal (eds.), Speech Coding and Synthesis, 495–518, Amsterdam, Lausanne, New York, Oxford, Shannon, Tokyo, Elsevier (1995).

[8] Akagi, M. and Ienaga, T., "Speaker individuality in fundamental frequency contours and its control," J. Acoust. Soc. Jpn(E), 18(2), 73–80 (1997).

[9] Dang, J. and Honda, K., "Acoustic characteristics of the piriform fossa in models and humans," J. Acoust. Soc. Am., **101**, 456–465 (1997).

[10] Takemoto, H., Kitamura, T., Honda, K., and Masaki, S., "Deformation of the hypopharyngeal cavities due to F0 changes and its acoustic effects," Acoust. Sci. & Tech. (accepted).