 

ACOUSTICAL LETTER

# Auditory discrimination of natural speech and synthetic speech used as voice disguise

Kanae Amino*, Hisanori Makinae and Toshiaki Kamada

*National Research Institute of Police Science,*
*6–3–1, Kashiwanoha, Kashiwa, 277–0882 Japan*

(*Received 31 July 2017, Accepted for publication 8 September 2017*)

**Keywords:** Voice forgery, Forensics, Concatenative speech synthesis, Training in phonetics
**PACS number:** 43.72.Uv [doi:10.1250/ast.39.48]

## 1. Introduction

Voice disguise is a challenging obstacle in practical speech investigations not only in biometrics but also in forensics. Impostors may change phonation by using glottal fry or whisper voice; change pitch; change accents by using a foreign accent or another dialect; or do anything else in their power to hide their identities. Especially in forensics, the importance of awareness of voice disguise has been emphasised since the 1970s [1–10]. In the last two decades, researchers have warned of a new type of voice disguise, the use of synthetic speech [11–13]. Synthetic speech was not a big problem for forensics when it first appeared, since it was easily distinguished from natural speech. However, as the quality of synthetic speech has steadily improved and the applications for speech synthesis have become of more affordable prices, its usage has become more accessible and hence, it has become a threat to researchers working in biometric and forensic fields.

In biometrics, Masuko and colleagues [11] tested voice disguise using synthetic speech in an HMM (hidden Markov model)-based automatic speaker verification system. They examined various factors, including the amount of training data and whether dynamic features were used in the speech synthesis, and concluded that in all cases they could not correctly reject the synthetic speech. Sato *et al.* [12] investigated methods for discriminating natural speech and HMM-based synthetic speech. They focused on the inter-frame variations of likelihood in speaker's GMM (Gaussian mixture model), and they succeeded in drastically reducing the false acceptance rate of synthetic speech under their experimental conditions. Nevertheless, they pointed out the possibility of voice forgery using the synthetic speech created by the concatenative method. Galou [13] confirmed that an HMM-based synthetic speech utterance could fool a state-of-the-art speaker recognition system. He also warned that the detection of synthetic speech would be mandatory in forensics.

Unlike biometrics, where automatic recognition is predominant, auditory detection of voice disguise has been attracting the forensic researchers' attentions, because forensic speech investigations, including voice comparison, voice profiling, and speaker classification, are carried out by human experts. These experts are often phoneticians who have obtained additional training in forensic science, or engineers or computer scientists who have obtained additional training in phonetics and forensic science [14]. World-wide surveys on forensic speech investigation [15,16] show that the most popular approach employed in most countries are auditory analysis, together with acoustic-phonetic analysis. Reich [4] conducted an experiment where two groups of listeners, naive (undergraduate students) and sophisticated (doctoral students and professors in speech and hearing sciences), discriminated disguised and undisguised speech created by forty male speakers. For the disguised speech, the speakers could freely select the disguise strategies. The average percentages of correct discrimination were 89.4 and 92.6 for naive and sophisticated listeners, respectively. Hirson and Duckworth [5] focused on glottal fry as voice disguise. They reported that the trained listeners could match speakers with 65% and 90% accuracy for disguised and undisguised speech, respectively.

When these studies were conducted in the 1980s and 1990s, synthetic speech may not have been a candidate of voice disguise. However, as mentioned above, the quality of synthetic speech has improved drastically; and it is a problem that there is no study that investigated human discrimination of natural speech and synthetic speech used as voice disguise. We should have some knowledge on whether synthetic speech can fool not only speaker recognition systems but also human ears. In this study, we investigated auditory discrimination of synthetic and natural speech. We also examined whether a knowledge of phonetics and speech synthesis is helpful in this discrimination.

## 2. Methods

### 2.1. Speech materials

Natural speech utterances of ten speakers (five female, S1–S5, and five male, S6–S10, age ranged from 19–26, with an average of 22.1 years old) were recorded. The speakers were all native speakers of Japanese and had normal hearing. The utterances were eight short Japanese sentences consisting of two clauses. The lengths of the sentences ranged from five to nine morae (7.25 morae on average). Recordings were made using a microphone (SONY ECM-23F5) and a PCM recorder (Marantz PMD671) in the anechoic room at the National Research Institute of Police Science (NRIPS). The speech data were sampled at 44.1 kHz with 16 bit resolution, then down-sampled at 16 kHz before the experiment in co-ordination with the specifications of the synthetic speech.

*e-mail: amino@nrips.go.jp

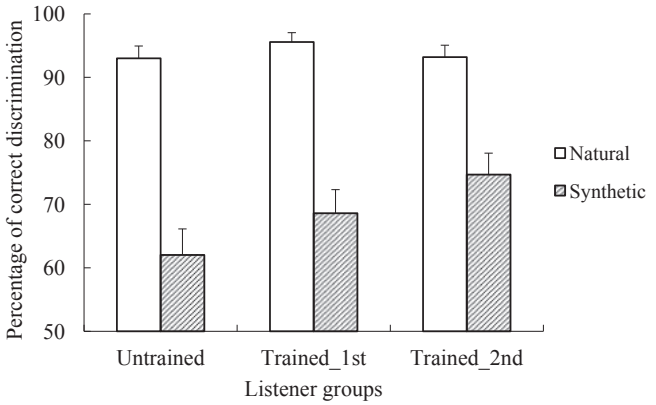


**Fig. 1** Results of the experiments.

Speech utterances of eight characters (C1–C8) were synthesised by using three different speech synthesis applications (A1–A3), which were on the market. According to the catalogues, all three applications employ the concatenative synthesis method. Each application can produce one or more characters' speech. The characters C1 to C5 were synthesised by A1, C6 by A2, and C7 and C8 by A3. The utterances were the same eight sentences as the natural speech utterances created by the above speakers. The synthesised speech data were sampled at 16 kHz with 16 bit resolution.

2.2. Procedures

Participants consisted of twenty native speakers of Japanese. Ten of them had no training and the rest of them had had an experience of university-level phonetic training. The trained listeners had taken a course in phonetics or linguistics including phonetics at universities. The experiment was conducted in a sound-treated room at NRIPS.

First, all of the participants underwent a hearing screening. Then they took a natural-synthetic discrimination test. The test consisted of 288 trials, 144 stimuli (eight sentences each uttered by the eighteen speakers/characters) with two repetitions, which were randomly presented on a computer using Praat [17]. The participants were instructed to listen to each stimulus carefully through the headphones (Sennheiser HD650) and to choose whether they thought it was natural or synthetic. They were allowed to listen to each stimulus only once. The experiment for the untrained listeners finished here. After the first discrimination test, the trained listeners were given an explanation about the basic methods of concatenative speech synthesis, and then took the second discrimination test, which was the same as the first test. The explanation was made orally, without using any visual materials, and included the following: (1) the system uses a recorded corpus of natural speech; (2) the method is based on the concatenation of the segments; and (3) prosody is modelled independently and added to the concatenated speech signals.

## 3. Results and discussion

The results are summarised in Fig. 1. Average percentage of correct discrimination for all listeners' first test was 94.3 for natural speech, which was significantly higher than the same result for the synthetic speech, that is, 65.3 ($t(19) = -5.12$, $p < 0.001$). The effect of the speakers/characters ($F(17, 136) = 40.39$, $p < 0.001$) was significant; however, that of the sentences was not ($F(7, 136) = 0.27$, $p = 0.97$). Average percent correct discrimination for each speaker/character is shown in Table 1. We can see that the synthetic male character C4 obtained the lowest score. It was significantly lower than the scores of the other four characters that had been synthesised by the same application. This suggests that the ease of natural-synthetic discrimination depends on the original speaker whose recorded data are used to create the synthesis. Among the natural speech, the speaker S4 obtained a remarkably low score. In the questionnaire, we found that this speaker had had voice training to be an announcer. The narrative pronunciation of announcers may be stereotypical of synthetic speech, which is created by using the speech uttered by professional speakers, and this may have had an effect on the discrimination success rate.

**Table 1** Average percentage of correct discrimination for speakers/characters.

| | Speaker/ Character | Application (synthetic speech) | Properties | Average %correct |
|---|---|---|---|---|
| Natural | S1 | — | Female | 97.5 |
| | S2 | | | 98.4 |
| | S3 | | | 91.6 |
| | S4 | | | 74.4 |
| | S5 | | | 96.9 |
| | S6 | | Male | 96.3 |
| | S7 | | | 96.6 |
| | S8 | | | 98.4 |
| | S9 | | | 95.9 |
| | S10 | | | 96.9 |
| Synthetic | C1 | A1 | Female | 79.4 |
| | C2 | | | 65.9 |
| | C3 | | Male | 73.1 |
| | C4 | | | 31.6 |
| | C5 | | Girl | 57.8 |
| | C6 | A2 | Female | 84.1 |
| | C7 | A3 | Girl | 62.2 |
| | C8 | | Boy | 68.4 |

A two-way ANOVA on the stimuli (natural or synthetic) of the first test and the listener groups revealed that the main effect of the stimuli ($F(1, 284) = 281.58$, $p < 0.001$) and the listener groups ($F(1, 284) = 6.40$, $p < 0.05$) were both significant, but their interaction was not significant ($F(1, 284) = 1.34$, $p = 0.25$). The scores for natural speech were better than those for synthetic speech for both untrained and trained listeners. The trained listeners performed better than the untrained listeners on both synthetic and natural

speech stimuli. This implies that a knowledge of phonetics is helpful for the discrimination task.

In comparing the trained listeners' first and second tests in Fig. 1, we can see that their performance on synthetic speech improved, whereas that on natural speech was slightly degraded. A two-way ANOVA on the stimuli and the tests showed that the main effect of the stimuli ($F(1, 284) = 203.75$, $p < 0.001$) and the interaction ($F(1, 284) = 7.07$, $p < 0.01$) were significant. The results of a post-hoc test revealed that the difference between natural and synthetic stimuli was significant ($p < 0.01$), whereas that between the first and second tests was not significant. A knowledge of speech synthesis may be helpful for identifying the synthetic speech; however its effect is not statistically significant.

## 4. Conclusions

In this study, we investigated the discrimination between natural and synthetic speech by untrained and phonetically trained listeners. We found that the listeners all tended to misperceive synthetic speech as natural speech, but not the other way around. The results also showed that a knowledge of phonetics is helpful for the discrimination; however, the effect of a knowledge of speech synthesis may be limited.

In forensics, speech investigators should keep in mind the possibility that the target speech materials may be synthetic before they start the analyses. In addition, they had better obtain a training in phonetics, if her/his academic discipline at university is not phonetics. Further research on the perception of the trained listeners will make clear what speech features forensic speech investigators should focus on, and it will also help us decide the details of the training in phonetics.

## Acknowledgments

Part of this study was presented at the 21st Annual Meeting of the Japanese Association of Forensic Science and Technology in 2015. The procedures used in this study were in accordance with the ethical standards of human experimentation proposed by the National Research Institute of Police Science and with the Helsinki Declaration. This work was partly supported by Grants-in-Aid for Scientific Research (24710195, 26350466, 26870865) from MEXT.

## References

[1] W. Endres, W. Bambach and G. Floesser, "Voice spectrograms as a function of age, voice disguise, and voice imitation," *J. Acoust. Soc. Am.*, **49**, 1842–1848 (1971).
[2] A. R. Reich, K. L. Moll and J. F. Curtis, "Effects of selected vocal disguises upon spectrographic speaker identification," *J. Acoust. Soc. Am.*, **60**, 919–925 (1976).
[3] A. R. Reich and J. E. Duke, "Effects of selected vocal disguises upon speaker identification by listening," *J. Acoust. Soc. Am.*, **66**, 1023–1028 (1979).
[4] A. R. Reich, "Detecting the presence of vocal disguise in the male voice," *J. Acoust. Soc. Am.*, **69**, 1458–1461 (1981).
[5] A. Hirson and M. Duckworth, "Glottal fry and voice disguise: A case study in forensic phonetics," *J. Biomed. Eng.*, **15**, 193–200 (1993).
[6] H. J. Künzel, J. Gonzalez-Rodriguez and J. Ortega-García, "Effect of voice disguise on the performance of a forensic automatic speaker recognition system," *Proc. Speaker and Language Recognition Workshop*, 4 pages (2004).
[7] P. Perrot, C. Preteux, S. Vasseur and G. Chollet, "Detection and recognition of voice disguise," *Proc. Int. Assoc. Forensic Phonetics Acoust.*, 3 pages (2007).
[8] C. Zhang and T. Tan, "Voice disguise and automatic speaker recognition," *Forensic Sci. Int.*, **175**, 118–122 (2008).
[9] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre and P. De Leon, "Speaker recognition anti-spoofing," in *Handbook of Biometric Anti-Spoofing*, S. Marcel, M. S. Nixon and S. Z. Li, Eds. (Springer-Verlag, London, 2014), pp. 125–146.
[10] H. Hollien, G. Didla, J. D. Harnsberger and K. A. Hollien, "The case for aural perceptual speaker identification," *Forensic Sci. Int.*, **269**, 8–20 (2016).
[11] T. Masuko, K. Tokuda and T. Kobayashi, "Imposture against a speaker verification system using synthetic speech," *IEICE Trans.*, **J83-D-II**, 2283–2290 (2000) (in Japanese).
[12] T. Sato, T. Masuko, T. Kobayashi and K. Tokuda, "Discrimination of synthetic speech generated by an HMM-based speech synthesis system for speaker verification," *IPSJ Trans.*, **43**, 2197–2204 (2002) (in Japanese).
[13] G. Galou, "Synthetic voice forgery in the forensic context: A short tutorial," *Forensic Speech and Audio Analysis Working Group*, pp. 1–3 (2011).
[14] M. Jessen, "Forensic phonetics," *Lang. Linguist. Compass*, **2**, 671–711 (2008).
[15] E. Gold and P. French, "International practices in forensic speaker comparison," *Int. J. Speech Lang. Law*, **18**, 293–307 (2011).
[16] G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs and C. G. Dorny, "INTERPOL survey of the use of speaker identification by law enforcement agencies," *Forensic. Sci. Int.*, **263**, 92–100 (2016).
[17] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, **5**, 341–345 (2001).

**Kanae Amino** received the B.A. and M.A. degrees in linguistics and Ph.D. degree in engineering from Sophia University, Tokyo, Japan, in 2002, 2004, and 2009, respectively. She was a post-doctoral researcher at Sophia University in 2009–2010 and a JSPS Research Fellow (PD) in 2010–2012. Currently she is a researcher at National Research Institute of Police Science (NRIPS) in Chiba, Japan. She has been working on the variations in speech production and the perception of the speaker individualities. Her research interests also include Japanese phonology and phonetics, and dialect and language identification for forensic purposes.

**Hisanori Makinae** received the B.E., M.E. and Ph.D. degrees in Electrical and Communication Engineering from Tohoku University, Miyagi, Japan, in 1998, 2000 and 2008, respectively. Since 2005, he has been with NRIPS, Chiba, Japan, where he is currently a senior researcher of Third Information Science Section. His research interests include speaker recognition in forensic field. He is a member of ASJ and the Japanese Association of Forensic Science and Technology (JAFST).

**Toshiaki Kamada** received the B.E. degrees in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1995. He has been with NRIPS, Chiba, Japan, since 1995. His research interests include speaker recognition and speech processing in forensic field. He is a member of ASJ and JAFST.