

ABSTRACT

We intuitively know that different people talk/write differently, even when they try to convey the same message. We also know that people tend to use their individually selected preferred words despite the fact that in principle they can use any word at any time from the vocabulary built up over their lives. This is due to the idiosyncratic choice of words, expressions and so forth.

In this study, we investigate the different use of fillers amongst Japanese speakers. More precisely speaking, we attempt to answer the question ‘to what extent we are idiosyncratic’ by empirically testing how well we can correctly identify same speakers as same speakers and recognise different speakers as different speakers purely based on the individual’s preferred selection of fillers in Japanese speech.

1 Introduction

Every speaker of a given language has their own distinctive and individual version of the language—which is often referred to as *idiolect* [1, 2]. This idiolect manifests itself in various aspects of communication, such as the choice of words, expressions, or even grammar, morphology, semantics and discourse structure.

Many factors contribute to the composition of this idiolect, such as gender [3], generation, dialect, personality [4] and so on, or pure idiosyncrasy [5]. The idiosyncratic nature of word selection between speakers/writers has been studied in different fields, for example, to understand speaking styles of political leaders [6], to identify the authors of literary works [7], to detect plagiarism [8] and to enhance the performance of automatic speaker recognition [9].

The current study investigates ‘to what extent we are idiosyncratic’ (in other words, how much we are different) in selecting certain words rather than others. We focus on the use of fillers in Japanese speech in this study as several studies impressionistically report some preferred choices of fillers in Japanese amongst different speakers [10, 11, 12]. We attempt to answer the question ‘to what extent we are idiosyncratic’ by investigating how well we can correctly identify same speakers as same speakers and recognise different speakers as different speakers purely based on the individual’s preferred selection of fillers.

2 Methodology

We selected non-contemporaneous speech samples (262 speakers x 2 sessions = 524 speech samples) from the Corpus of Spontaneous Japanese (CSJ) [13], and compared the performance of a speaker recognition system by setting different parameters. These 262 speakers consist of 163 male speakers and 99 female speakers. The detailed explanation of the CSJ is given in §2.1.

Two kinds of comparisons are involved in speaker recognition tests. One is called *Same Speaker Comparison* where two speech samples produced by the same speaker need to be correctly identified that they are from the same speaker and *mutatis mutandis* the other is *Different Speaker Comparison*.

In Experiment 1, all 262 speakers are used to conduct a series of experiments. In Experiment 2, a series of experiments are conducted separately for male and female speakers under more constrained conditions. The detailed procedures of Experiments 1 and 2 are explained in §3 and §4, respectively.

2.1 Database and speakers

- The Corpus of Spontaneous Japanese (CSJ) [13] contains a large number of monologues, either Academic Presentation Speech (APS) or Simulated Public Speech (SPS).
- APS was mainly recorded live at academic presentations, most of which were 12–25 minutes long. For SPS, 10–12 minute mock speeches on everyday topics were recorded.
- 262 speakers (163 male and 99 female speakers) who have two non-contemporaneous recordings (262 speakers x 2 sessions = 524 speech samples) were selected for this study.
- CSJ gives a five-scale rating on spontaneity of the speech (i.e. not reading).

2.2 Fillers

| Fillers | Count |
|---------|-------|
| e- | 27776 |
| e | 12046 |
| ma | 8816 |
| ano- | 7213 |
| ano | 6988 |
| ma- | 5990 |
| so | 2533 |
| e-to | 2479 |
| a | 2364 |
| n | 1924 |
| o- | 1559 |
| o | 1376 |
| e-to | 1327 |
| sono- | 1127 |
| u | 920 |

Table 2: Fillers.

In Wikipedia (cited on 24 April 2009), fillers are concisely defined as ‘... sounds or words that are spoken to fill up gaps in utterances ...’. In English, the most common filler sounds are *uh*, *er* and *um*.¹ However, different people use different definitions and terms for so-called *fillers*, resulting from different focuses in different disciplines. Like studies on fillers in English, a large number of research on fillers in Japanese can be found [12, 14, 15].

In CSJ, a filler tag is assigned to one of the pre-selected words given in Table 1 which have the function of ‘filling up gaps in utterances’. Some of the words given in Table 1 can be used as lexical words. If it is uncertain as to whether a given word is used as a lexical word or a filler, additional information is embedded in the tag indicating this uncertainty. In this case, the word was removed from the experiments.

From the 524 speech samples, 49 different fillers, out of which the 15 most frequent fillers are listed in Table 2, were identified.

- a(-), i(-), u(-), e(-), o(-), n(-), to(-)†, ma(-)†
- u(-)n, a(-)nno(-)†, so(-)nno(-)†
- u(-)n(-)to(-)†, a(-)to(-)†, e(-)to(-)†, n(-)to(-)†
- one of the above + {~desune(-), ~ssune(-)}
- one of the above with † + {ne(-), sa(-)}

Table 1: Pre-selected fillers in CSJ.

2.3 Vector Space Model

Using the frequency counts of the identified fillers, each speech is represented as a real-valued vector in this study. If n different fillers are used to represent a given speech S , the dimensionality of the vector is n . That is, S is represented as a vector of n dimensions ($\vec{S} = \{F_1, F_2 \dots F_n\}$, where F_i represents the i^{th} component of \vec{S} and F_i is the frequency of the i^{th} filler). For example, if five fillers are used to represent a speech (X), and the frequency counts of these fillers are 38, 10, 4, 0 and 0 respectively, the speech X is represented as $\vec{X} = \{38, 10, 4, 0, 0\}$.

2.4 Term Frequency Inverse Document Frequency Weighting

If a given word is frequently used by many speakers, it is apparent that this word is not as useful for speaker recognition tasks as those words which are used by a limited number of speakers. In this study, different weights are given to different fillers depending on the significance of a filler for speaker recognition. The *tf-idf* (term frequency inverse document frequency) weight (Formula 1) is used to evaluate how important a filler is to a speech in a collection, and a weight is given to a filler according to its importance.

$$W_{i,j} = tf_{i,j} * \log\left(\frac{N}{df_i}\right) \quad (1)$$

| | f_1 | f_2 | f_3 | f_4 |
|-------|------------|------------|------------|------------|
| s_1 | 1 (0.9163) | 0 (0) | 2 (0.4463) | 1 (0.5108) |
| s_2 | 3 (2.7489) | 2 (0.4463) | 1 (0.2231) | 0 (0) |
| s_3 | 0 (0) | 1 (0.2231) | 0 (0) | 2 (1.0217) |
| s_4 | 0 (0) | 3 (0.6694) | 1 (0.4463) | 1 (0.5108) |
| s_5 | 0 (0) | 2 (0.4463) | 3 (0.6694) | 0 (0) |
| df | 2 | 4 | 4 | 3 |

Table 3: An example matrix of fillers in speeches.

f_1 in s_1 ($tf = 1$) as an example, f_1 appears in two speech samples ($df = 2$) and the number of speech samples is 5 ($N = 5$). Therefore, the *tf-idf* weighted value of the frequency of f_1 in s_1 is $0.9163 (= 1 * \log(\frac{5}{2}))$.

2.5 Cosine Similarity Measure

The difference between two speech samples, which are represented as vectors (\vec{x}, \vec{y}), is calculated based on the cosine similarity measure (Formula 2) in this study. In Formula 2, the value of $diff(\vec{x}, \vec{y})$ becomes smaller, the greater the difference between two vectors (speech samples) is.

$$diff(\vec{x}, \vec{y}) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|} = \frac{\sum_{i=1}^n x_i * y_i}{\sqrt{\sum_{i=1}^n x_i^2} * \sqrt{\sum_{i=1}^n y_i^2}} \quad (2)$$

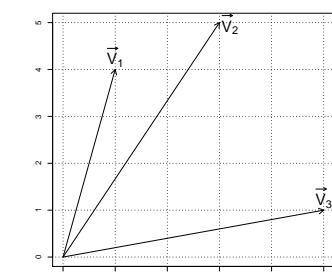


Fig. 1: Three 2-dimensional vectors.

2.6 Speaker Recognition Performance

In this study, the performance of a speaker recognition system is assessed on the basis of the probability distributions of the difference for two contrastive hypotheses. One is the hypothesis that two speech samples were uttered by the same speaker (the same speaker hypothesis = the SS hypothesis) and the other is that two speech samples were uttered by different speakers (the different speaker hypothesis = the DS hypothesis). These probabilities can be formulated as $P(E|H_{ss})$ and $P(E|H_{ds})$ respectively, where E is the difference, H_{ss} is the SS hypothesis and H_{ds} is the DS hypothesis. The ratio of these two probabilities (or likelihood) is called likelihood ratio (LR) ($= \frac{P(E|H_{ss})}{P(E|H_{ds})}$). LR tells us how much more likely one hypothesis is than the other for a given piece of information. A detailed example is given in §3 regarding the assessment of a speaker recognition system.

3 Experiment 1

In Experiment 1, 524 speech samples collected from all of the 262 speakers are used for speaker recognition tasks. In Figure 2, the probability distributions of the difference of two speech samples are plotted separately for the SS and DS hypotheses, respectively ($P(E|H_{ss})$ and $P(E|H_{ds})$). Since 262 speakers are involved in Experiment 1, 262 same speaker comparisons and 136764 different speaker comparisons are possible.

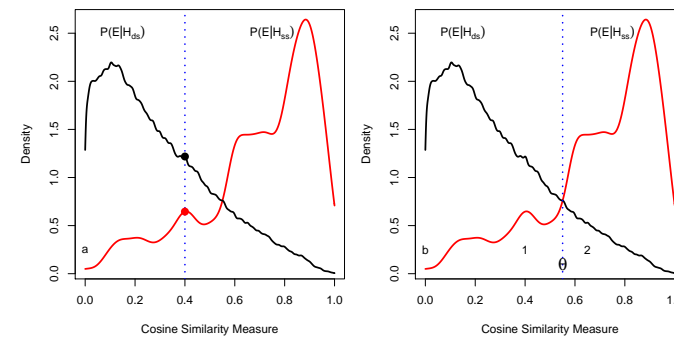


Fig. 2: The probability distributions of the difference between two speech samples for the same speaker and different speaker hypotheses.

Suppose that you have two speech samples, and that you do not know whether they are from the same speaker or different speakers. Using the distributions of the two conditional probabilities given in Figure 2, you can predict whether the difference observed between these speech samples is more likely to occur if they are from the same speaker or different speakers. For example, one calculates that the difference between these two speech samples is 0.4 (which is the blue dotted vertical line of Figure 2a). According to the probability density distributions of $P(E|H_{ss})$ and $P(E|H_{ds})$, $P(0.4|H_{ss}) = 0.645417$ and $P(0.4|H_{ds}) = 1.128122$. That is, one can say that the difference of 0.4 is more likely to be obtained from different speakers than from the same speaker. If a threshold (θ) is set at the crossing point of the two distributions (which is the vertical dotted line of Figure 2b), it can be said that the difference between two speech samples is more likely to be from the same speaker if the difference is higher than θ and from different speakers if the difference is lower than θ .

These two distributions given in Figure 2b can also tell the accuracy of the system. Area 1 in Figure 2b—the area surrounded by the red line, the vertical line and the line of $y = 0$ —is the predicted error for the same speaker comparisons, and Area 2 of Figure 2b—the area which is surrounded by the black line, the vertical line and the line of $y = 0$ —is the predicted error for the different speaker comparisons (Formula 3). By calculating the proportion of the errors against all same speaker or different speaker comparisons, it is possible to estimate the accuracy of the system.

$$ERROR_{ss} = \int_0^\theta f_{ss}(x)dx, ERROR_{ds} = \int_\theta^1 f_{ds}(x)dx \quad (3)$$

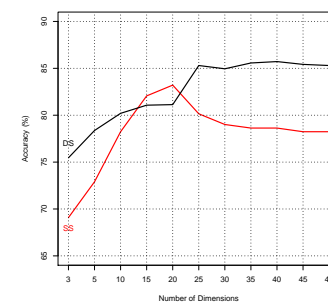


Fig. 3: The performance of the speaker recognition system as a function of the number of dimensions in spacial vectors.

4 Experiment 2

| Male | | Female | |
|---------|-------|---------|-------|
| Fillers | Count | Fillers | Count |
| e- | 16675 | e- | 4987 |
| ma | 5703 | ano- | 3129 |
| e | 5589 | ano | 2646 |
| ma- | 4128 | e | 2484 |
| ano | 3373 | ma | 1703 |
| ano- | 3255 | ma- | 1123 |
| sono | 1741 | sono | 564 |
| e-to | 1669 | n | 505 |
| a | 1325 | a | 489 |
| o- | 1102 | e-to | 447 |
| n | 1093 | sono- | 267 |
| a | 1059 | e-to | 251 |
| e-tto | 836 | n- | 233 |
| sono- | 788 | a- | 211 |
| u | 634 | etto | 182 |

Table 4: Fillers for male and female.

speakers in terms of frequency, it can be seen from Table 4 that very similar fillers are used as the 15 most frequent fillers amongst male and female speakers.

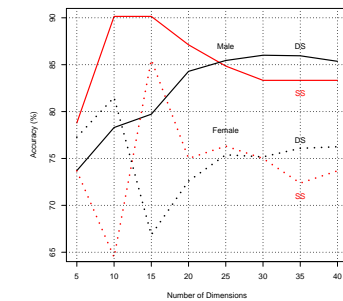


Fig. 4: The performance of the speaker recognition system as a function of the number of dimensions in spacial vectors, plotted separately for male and female speakers.

when 20 or more fillers are included in the vectors, and 2) the speaker recognition for females outperforms that for males by approximately 10%.

5 Discussion and conclusions

The results of Experiment 1 demonstrate that the accuracy of same speaker recognition and that of different speaker recognition can be higher than 80%. That is, 1) the fillers carry the idiosyncratic information of speakers in Japanese to the extent that more than 80% of different speakers can be correctly identified as different, and 2) the idiosyncrasy of each speaker with respect to fillers is consistent to the extent that more than 80% of same speakers can be correctly identified as the same. In other words, we are diverse in the use of fillers in Japanese to the degree that more than 80% of different speakers can be correctly identified as different.

The results of Experiment 2 indicate that there are some differences in terms of the degree of within-speaker variability in the use of fillers between male and female speakers in that female speakers are more varied in the use of fillers across different speeches than male speakers. That is, male speakers are more consistent with their selection of fillers than female speakers.

6 Acknowledgements

This study was financially supported by the College of Asia and the Pacific, the Australian National University.

References

- [1] M. A. K. Halliday, A. McIntosh and P. Stevens. 1964. *The Linguistic Sciences and Language Teaching*. London: Longmans.
- [2] M. Coulthard and A. Johnson. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London, New York: Routledge.
- [3] M. Koppel, S. Argamon and A. R. Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.
- [4] J. W. Pennebaker and L. A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.
- [5] A. Peters and L. Menn. 1993. False starts and filler syllables: Ways to learn grammatical morphemes. *Language*, 69:742–777.
- [6] R. Slatcher, C. Chunga, J. Pennebaker and L. Stone. 2004. Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality*, 41:63–75.
- [7] R. Thisted and B. Efron. 1987. Did Shakespeare write a newly-discovered poem?. *Miometrka*, 77:445–455.
- [8] D. Woolls. 2003. Better tools for the trade and how to use them. *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 10(1):102–112.
- [9] G. Doddington. 2001. Speaker recognition based on idiolectal differences between speakers, *Proceedings of the Eurospeech*, Aalborg, Denmark, September 2001.
- [10] S. Furui, K. Maekawa and H. Isahara. 2002. Intermediate results and perspectives of the project ‘Spontaneous Speech: Corpus and Processing Technology’, *Proceedings of the 2nd Spontaneous Speech Science and Technology Workshop*, 1–6.
- [11] Y. Sato. 2002. ‘UN’ and ‘SO’ in Japanese Casual Conversation between Native Speakers: *The Use of Fillers*, MA thesis, Nagoya University.
- [12] T. Yamane. 2002. *Fillers in Japanese Conversation*, Tokyo: Kuroshio publisher.
- [13] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. 2000. Spontaneous speech corpus of Japanese, *The Second International Conference of Language Resources and Evaluation (LREC2000)*, Athens, 2000, 947–952.
- [14] T. Nagura. 1997. Hesitations in Japanese. *Sekaino Nohongo Kyoouku*, 7:201–218.
- [15] M. K. Philips. 1998. *Discourse markers in Japanese Connectives, fillers and inter-actioal particles*, PhD dissertation, Michigan State University.