Filler Words as a Speaker Classification Feature

Shunichi Ishihara[1], *Yuko Kinoshita*[2]

[1]School of History, Culture and Language, the Australian National University, Australia
[2]Faculty of Arts and Design, University of Canberra, Australia
shunichi.ishihara@anu.edu.au, yuko.kinoshita@canberra.edu.au

Abstract

Many of us can think of iconic words or expressions that our close friends or family often use. This study investigated the potential of idiosyncrasies in usage of filler words as a speaker classification parameter, using spontaneous Japanese speech. We estimated likelihood ratios (LRs) based on which filler words were used and how often, and performed speaker discrimination tests using the LRs. As a result we discovered that speakers' choices of fillers can predict speaker's identity to some degree, and that this has potential as an additional feature for speaker classification.

Index Terms: fillers, idiosyncrasy, likelihood ratio, speaker classification, Japanese

1. Introduction

We often observe individual characteristics in the use of vocabularies. In our day–to–day speech, we tend to use a limited part of our vocabulary repeatedly. This phenomenon can be interpreted as an aspect of each person's own distinctive and individualised version of the language—an *idiolect* [1, 2]. This idiolect manifests itself in many aspects of communication, including use of words and expressions; or even grammar, morphology, semantics and discourse structure. The idiosyncratic nature of word selection by speakers (and writers in written communication) has been studied from various perspectives, including: analysis of speaking styles of political leaders [3], identification of the authors of literary works [4], detection of plagiarism [5] and enhancing the performance of automatic speaker recognition [6].

So forensic linguists ask: how can we use the idiolect concept in speaker classification? Idiolects define speaker–to–speaker variations in the use of the language, and 'speaker–to–speaker variation' is a key concept in speaker classification. This preliminary study focuses on a particular aspect of a specific group of idiolects: variations in the use of filler words in Japanese. Filler words are defined as the words that 'fill up gaps in utterances'. This definition was employed to identify fillers in the database used in this study. Several preceding studies observed that there are some speaker dependent variations in the selection of those filler words [7, 8, 9, 10]. Yet, many of these studies are based on subjective observations.

The non–acoustic feature used for the current study is very robust against unfavourable recording conditions. Unlike features such as MFCC, the effects that the channel difference or background noise can have on this feature is minimal.

We explore two research questions in this study: 1) is choice of filler words idiosyncratic, and 2) if it is, how useful is this idiosyncrasy in speaker classification. In order to answer these questions, we parametrised the selection of filler words, performed speaker discrimination tests using the likelihood ratios (LRs), and tested the performance of this feature using equal error rate (EER) and C_{llr} calculations.

2. Methodology

2.1. Database and speakers

For the data, we used the Corpus of Spontaneous Japanese (CSJ) [11], which contains the recordings of various speaking styles such as sentence reading, monologue, and conversation. For this study we used only the monologues, categorised as 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. We selected our speakers from this corpus based on three criteria: availability of multiple and non–contemporaneous recordings, naturalness of the speech, and speaking in standard modern Japanese. Naturalness and standardness of the language was assessed on the basis of the rating the CSJ provides. This gave us 264 speech samples (or monologues): 132 male speakers, each in 2 sessions. Keeping an application to forensic speaker classification in mind, we selected only male speakers.

2.2. Extraction of filler words

In CSJ, a filler tag is assigned to the pre-selected words given below which have the function of filling in gaps in utterances. A parenthesis indicates an optional segment and "-" stands for the prolongation of the immediately preceding segment. Some of these words can be used as lexical words as well as fillers, and this distinction is sometimes difficult to make. Where these ambiguous uses of words were tagged in CSJ, they were excluded from our data.

- a(-), i(-), u(-), e(-), o(-), n(-), to(-)†, ma(-)†
- u(-)n, a(-)(n)no(-)†, so(-)(n)no(-)†
- u(-)n(-)(t)to(-)†, a(-)(t)to(-)†, e(-)(t)to(-)†, n(-)(t)to(-)†
- one of the above + {desune(-), ssune(-)}
- one of the above with † + {ne(-), sa(-)}

In the selected 264 speech samples, we observed 44 different filler words, which are listed in Table 1.

2.3. Parametrisation

In order for the choice of the filler words to be useful as a speaker classifier, it has to satisfy two criteria. First it has to be consistent within a speaker. If a speaker uses a certain word consistently and frequently, this particular filler word starts to have some significance in characterising the speaker. The second criterion is the uniqueness of the use. Assume that one speaker uses a certain filler very frequently. This by itself does

ISBN 978-0-9581946-3-1
Accepted after peer review of full paper

14 – 16 December 2010, Melbourne, Australia

Table 1: *Fillers and their frequencies of occurrences.*

Fillers	Count	Fillers	Count	Fillers	Count
e-	16675	u-	571	nto	16
ma	5703	n-	526	to-	16
e	5589	o	437	nto-	15
ma-	4128	etto	392	u-n	9
ano	3373	i-	313	a-to	5
ano-	3255	e-to-	301	n-to-	4
sono	1741	i	186	nto-	4
e-to	1669	eto	160	a-to-	3
a	1325	etto-	131	ntto	1
o-	1102	to	118	n-tto-	1
n	1093	e-tto-	114	u-nto	1
a-	1059	a-no-	111	a-tto	1
e-tto	836	a-no	100	e-ttodesune	1
sono-	788	un	51	so-no-	1
u	634	eto-	29		

not give us much information unless we know how often others would use this particular word. If everyone else also uses it often, then this word will not give us much information on the speaker's identity. However, if it is a particularly unusual word and no one else uses it, the frequent use of this word can provide us with some useful information. In order to capture those characteristics, we have to parametrise the information presented in Table 1. We describe our method below.

2.3.1. *Vector space model of filler data*

Using the occurrence counts of the identified fillers, each speech is modelled 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 ($\overrightarrow{S} = \{F_1, F_2 \dots F_n\}$, where F_i represents the i^{th} component of $\overrightarrow{S}$ and F_i is the occurrence count of the i^{th} filler). For example, if five fillers are used to represent the use of filler words in speech X, and the occurrence counts of these fillers are 3, 10, 4, 18 and 1 respectively, the filler information in the speech X is represented as $\overrightarrow{X} = \{3, 10, 4, 18, 1\}$.

2.3.2. *Weighting for uniqueness*

The usefulness of particular words is determined by their uniqueness as well as how frequently they are used. Thus, different weights were given to different filler words depending on their uniqueness in the pooled data. The *tf* · *idf* (term frequency inverse document frequency) weight (Formula 1) is used to evaluate how unique a given filler word is in the population, and a weight is given to that filler to reflect its importance to the speaker discrimination [12]. In Formula 1, term frequency ($tf_{i,j}$) is the number of occurrences of word *i* (W_i) in the speech sample *j* (d_j). Document frequency (df_i) is the number of speech samples in the pooled speech data in which that word *i* (W_i) occurs. *N* is the total number of speech samples.

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

2.3.3. *Cosine similarity measure*

The difference between two speech samples, which are represented as vectors ($\overrightarrow{x}$, $\overrightarrow{y}$), is calculated based on the cosine similarity measure (Formula 2) [12]. This particular method was

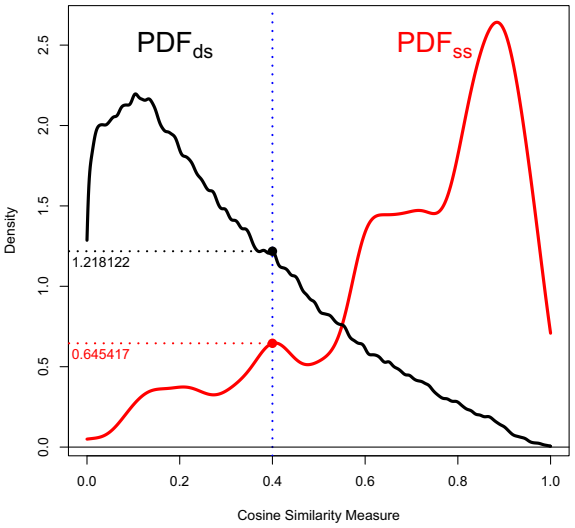

Figure 1: *Examples of PDFss and PDFds.*

selected as the vectors of each speech model varied in length.

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

The range of the difference in two vectors ($diff(\overrightarrow{x}, \overrightarrow{y})$) is between 1.0 (=$cos(0^\circ)$) for two vectors pointing in the same direction and 0.0 (=$cos(90^\circ)$) for two orthogonal vectors.

3. Calculation of Likelihood Ratios (LR)

An LR is the probability that the evidence would occur if an assertion is true, relative to the probability that the evidence would occur if the assertion is not true [13]. In the context of forensic voice comparison, it will be the probability of observing the difference between two speech samples if they had come from the same speaker (the SS hypothesis), relative to the probability of observing the same difference if it had been produced by different individuals (the DS hypothesis). Letting P represent probability, E evidence, H_{ss} the SS hypothesis and H_{ds} the DS hypothesis, LR can be expressed as $LR = \frac{p(E|H_{ss})}{p(E|H_{ds})}$. The LR will be larger than unity when the given evidence supports the SS hypothesis, and smaller than unity when the evidence supports the DS hypothesis. The relative distance of the given LR from unity quantifies the strength of the evidence.

In this study, we produced likelihood ratios (LRs) using those similarity scores. For the calculation, we first pooled the similarity scores for the same speaker (SS) comparisons and different speaker (DS) comparisons separately. Then each probability distribution function (PDF) was modelled using the kernel density function (`KernSmooth` library of R statistical package). Examples of an SS PDF (PDF_{ss}) and a DS PDF (PDF_{ds}) are given in Fig. 1.

We then calculated LRs based on these two PDFs using the cross–validation approach. Suppose that you obtained two speech samples of unknown origin. Using the methodology described above, you calculated the cosine similarity measure of these two speech samples and it was, say, 0.4 (which is the blue dotted vertical line of Fig. 1). From the density values

(y–axis) of the PDFs which correspond to the cosine similarity measure (x–axis) of 0.4, you can obtain an LR of 0.529 ($= \frac{p(0.4|H_{ss})}{p(0.4|H_{ds})} = \frac{0.645417}{1.218122}$). The LR of 0.529 indicates that the cosine similarity measure of 0.4 is more likely to be obtained from different speakers than the same speaker.

The calculation of LRs was repeated for differently–sized spatial vectors: 5, 10, 15, 20, 25, 30, 35, 40 and 44 dimensions. This is to see how the size of the vector dimension affects the performance of speaker classification and the LRs. The vectors were selected on the basis of their frequency of occurrence. For instance, the spatial vector of 5 dimensions means that the frequency counts of the five most frequently–used fillers are used to represent the speech sample.

4. Testing Results and Discussion

Using the obtained LRs, we performed speaker discrimination tests. Since the LRs greater than 1 support the SS hypothesis and those smaller than 1 support the DS hypothesis, we tested whether the LRs we obtained support the hypothesis which we know to be true.

We then evaluated our method using equal error rate (EER) and C_{llr} [14]. EER shows the error rate where the SS and DS comparisons achieve the same error rate. EER provides very useful information as to how accurately a given method can make binary decisions on the speakers' identities. However, in forensic contexts, it is necessary to examine the calibration of the LR. For example, although $\log_{10}$ LRs of 5 and 100 both support the same–speaker hypothesis, they indicate significantly different strengths of evidence. It is therefore extremely important to assess how well the LRs produced by each method are calibrated. C_{llr} is a metric that allows us to evaluate how well the obtained LRs are calibrated. C_{llr} was calculated using the `FoCal` toolkit [15]. C_{llr} can be expressed as the combination of C_{llr}^{min} (potential performance of the system when it is optimised) and C_{llr}^{cal} (amount of calibration needed to achieve the optimum performance). C_{llr} and C_{llr}^{cal} are calculated for each vector size and presented in Table 2 with its EER.

Table 2: *C_{llr}, C_{llr}^{cal}, EER and error % for SS and DS comparisons for cross–validated LRs with different dimensions in vectors.*

Dims.	C_{llr}	C_{llr}^{cal}	*EER*	SS Error %	DS Error %
44	0.530	0.028	0.167	16.67	13.92
40	0.530	0.028	0.167	16.67	14.31
35	0.530	0.028	0.167	17.42	13.72
30	0.528	0.030	0.164	16.67	13.58
25	0.520	0.041	0.153	15.90	14.05
20	0.496	0.040	0.154	14.39	15.46
15	0.533	0.072	0.201	10.60	19.98
10	0.569	0.055	0.176	10.60	21.48
5	0.785	0.083	0.278	28.79	26.18

Table 2 shows that performance improves notably as we increase the vector dimension size from 5 to 10. However, increase in the number of dimensions after this point does not improve the performance as dramatically, however. Particularly after the dimension size 20 (C_{llr} = 0.496), the number of dimensions does not appear to have any effect on the speaker discrimination performance.

We found that C_{llr}^{cal} is surprisingly small in all cases (0.083∼0.028). This suggests that the LRs produced using the selection of filler words generally reflect the strength of the evidence very well. Also, a constant improvement can be observed from the dimension of 5 to that of 30 with some minor ups and downs. After the dimension size reaches 30, C_{llr}^{cal} becomes static. This indicates that C_{llr}^{cal} improves even after the C_{llr} peaks at the dimension size 20. The fact that speaker classification performance peaks with half of the dimensions available is not surprising. The feature vectors were based on the occurrence counts of a given filler word, and we picked ones with higher frequency first to be included in the feature. So vectors in the later orders have very low frequencies, such as 0. This means that the latter part of longer vectors tends to include very similar low numbers across speakers. This would have introduced noise in the assessment of between-speaker difference, making them look more similar.

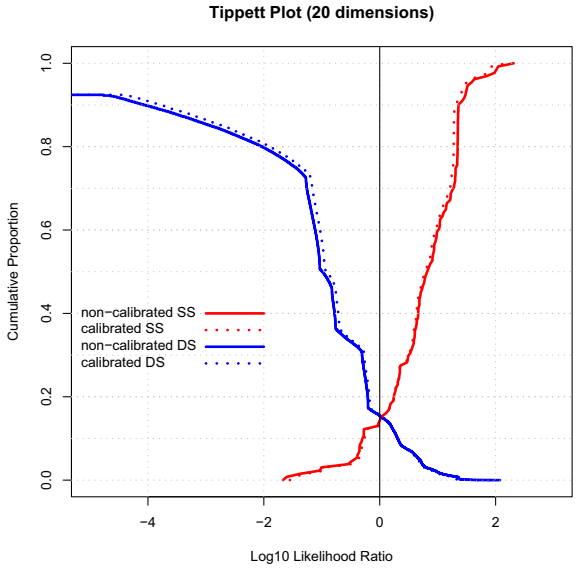

Figure 2: *Tippett plot of non–calibrated (solid lines) and calibrated (dotted lines) cross–validated LRs from the speaker discrimination system with vectors of 20 dimensions.*

A Tippett plot for the best performing system, dimension size 20, is presented in Fig. 2. In this figure, both calibrated (dotted line) and uncalibrated (solid line) log LRs are plotted. Fig. 2 shows that the cross point of two accumulative curves is in the vicinity of 0, where theoretically the threshold should fall for log LRs. We can also see that there is little difference between calibrated and uncalibrated LRs.

Furthermore, although the feature used here is based on speaker's selection of words, the LRs show some similarities to the ones produced using acoustic features, such as formant frequencies and F0 [16, 17, 18]. Comparing the LRs that correctly discriminated speaker pairs, we can see that the SS comparison produces less strong LRs than the DS comparisons. The SS pairs that showed the greatest strength of evidence were only about 200 times more likely to be observed, assuming SS provenance ($\log_{10}$ LR = 2.311), whereas the best DS trial has an astronomically low LR ($\log_{10}$ LR = –17.086). This may reflect the fact that we can have almost infinite ways to differ from other people, but the same speaker cannot be more similar beyond being identical.

However, we also observed some distinct differences between the feature used in this study and acoustic features [16, 17, 18]. The distributions of the LRs given in Fig. 2 for the SS and DS samples appear fairly symmetric (particularly between $\log_{10}$ LR of –2 and $\log_{10}$ LR of 2) even without calibra-

tion. For DS comparisons, the distribution of the LRs obtained from acoustic features is often much more spread out compared to its SS counterpart. It has been speculated that the asymmetric distribution of the LRs might be caused by differences in the testing size—generally, preceding research had a much larger number of DS comparisons than SS comparisons. However, the results we obtained here seem to oppose this theory. Interestingly, the distributions of the two types of LRs become even more similar as the dimension of the vectors increases. Fig. 3 presents LRs produced with a spatial vector of 44 dimensions. Here we can also observe how the inclusion of extra dimensions reduces the size of the LRs, especially in DS comparisons.

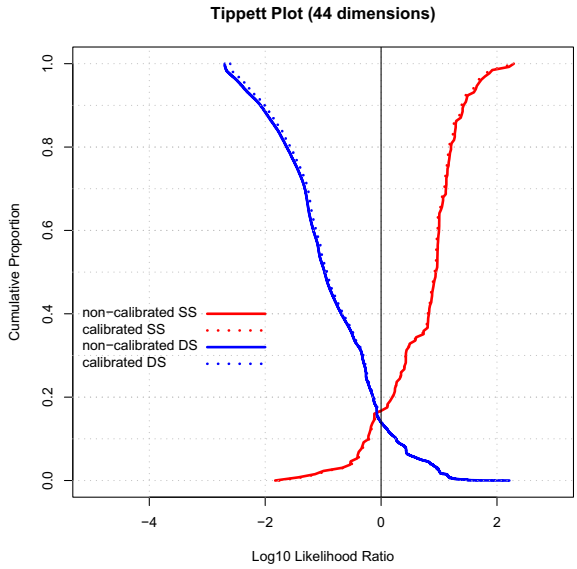

Figure 3: *Tippett plot of non–calibrated (solid lines) and calibrated (dotted lines) cross–validated LRs from the speaker discrimination system with vectors of 44 dimensions.*

5. Conclusions

In this study, we have demonstrated that there are individual characteristics in the use of Japanese filler words, and that it is indeed possible to classify speakers based on these individual characteristics at an EER of approximately 15%. EER 15% by itself is not a strong result, but this is based solely on one feature. Also, this feature is completely independent from the acoustic features normally used in speaker discrimination tasks. This feature thus has potential to make a significant contribution to improving the accuracy of speaker discrimination systems [19]. In future projects, we intend to extend this study by combining this feature with other, more conventional speaker discrimination features, such as formants, F0 or MFCC. The effect of the duration of speech also needs to be examined. We used relatively long speeches, so it is of interest how well this feature performs with shorter speeches. We also intend to test the same idea with languages other than Japanese.

6. Acknowledgements

This study was financially supported by the International Association for Forensic Phonetics and Acoustics and the College of Asia and the Pacific, the Australian National University. The authors thank anonymous reviewers for their valuable comments.

7. References

[1] M. A. K. Halliday, A. McIntosh and P. Strevens. 1964. *The Linguistic Sciences and Language Teaching*. London: Longman.

[2] M. Coulthard and A. Johnson. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London, New York: Routledge.

[3] R. Slatcher, C. Chunga, J. Pennebakera 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(1):63–75.

[4] R. Thisted and B. Efron. 1987. Did Shakespeare write a newly–discovered poem? *Biometrika*, 74(3):445–455.

[5] 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.

[6] G. Doddington. 2001. Speaker recognition based on idiolectal differences between speakers, *Proceedings of the Eurospeech*, Aalborg, Denmark, September 2001.

[7] 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.

[8] Y. Sato. 2002. *'UN' and 'SO' in Japanese Casual Conversation between Native Speakers: The Use of Fillers*. MA thesis, Nagoya University.

[9] C. Yamane. 2002. *Fillers in Japanese Conversation*. Tokyo: Kuroshio publisher.

[10] S. Ishihara. 2009. How diverse are we? An empirical analysis of individual differences in the use of fillers, *The 11th International Pragmatics Conference*, Unpublished paper.

[11] 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.

[12] C. D. Manning and H. Schütze. 2001. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

[13] B. Robertson and G. A. Vignaux. 1995. *Interpreting Evidence*. Chichester: Wiley.

[14] N. Brümmer and J. du Preez. 2006. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2–3):230–275.

[15] N. Brümmer. 2005. *FoCal Toolkit* [software], Available: http://www.dsp.sun.ac.za/ nbrummer/focal/

[16] J. Gonzalez–Rodriguez, P. Rose, D, Ramos, D. T. Toledano and J. Ortega–Garcia. 2007. Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, & Language Processing*, 15(7):2104–2115.

[17] G. Morrison. 2008. Forensic voice comparison using likelihood ratios based polynomial curves fitted to the formant trajectories of Australian English /ai/. *The International Journal of Speech, Language & the Law*, 15(2):249–266.

[18] Y. Kinoshita, S. Ishihara and P. Rose. 2009. Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. *The International Journal of Speech, Language & the Law*, 16(1):249–266.

[19] E. Shriberg and A. Stolcke. 2008. The case for automatic higher–level features in forensic speaker recognition, *Proceedings of Interspeech 2010*, 1509–1512.