Estimation of Speaking Style in Speech Corpora Focusing on speech transcriptions

Raymond Shen, Hideaki Kikuchi
Faculty of Human Sciences, Waseda Univiersity
359-1192, Mikajima 2-579-15, Tokorozawa, Saitama, Japan
E-mail: raymondshenrui@gmail.com, kikuchi@waseda.jp

Abstract

Recent developments in computer technology have allowed the construction and widespread application of large-scale speech corpora. To foster ease of data retrieval for people interested in utilising these speech corpora, we attempt to characterise speaking style across some of them. In this paper, we first introduce the 3 scales of speaking style proposed by Eskenazi in 1993. We then use morphological features extracted from speech transcriptions that have proven effective in style discrimination and author identification in the field of natural language processing to construct an estimation model of speaking style. More specifically, we randomly choose transcriptions from various speech corpora as text stimuli with which to conduct a rating experiment on speaking style perception; then, using the features extracted from those stimuli and the rating results, we construct an estimation model of speaking style by a multi-regression analysis. After the cross validation (leave-1-out), the results show that among the 3 scales of speaking style, the ratings of 2 scales can be estimated with high accuracies, which prove the effectiveness of our method in the estimation of speaking style.

Keywords: speaking style, transcriptions, estimation

1. Introduction

With the development of computing technologies and increasing needs of speech data, speech corpora are being constructed and several organizations that collect and manage linguistic resources have been grown. Linguistic Data Consortium, known as LDC (LDC), launched by University of Pennsylvania, USA and European Language Resources Association, also known as ELRA (ELRA), mainly being active in Europe could be raised as two leading organizations in this field. Searching services are also provided for users to select out suitable speech corpora to their intended purposes in huge quantities of resources. In Japan, we cooperated with NII-SRC (NII-SRC), which is also one of the organizations working on the distribution of speech corpora, and show the effectiveness of visualized searching systems based on attributes of corpora (Yamakawa, 2009)(Shen, 2011). However, besides attributes like "purpose" or "speakers", "speaking style" shall also be considered useful information. According to Jorden (Jorden, 1987), every language reflects stylistic differences, but it seems that these organizations above only give little information on speaking style (like dialogue, monologue), let alone on diversity in speaking style due to speakers or conditions even in a single speech corpus. To solve this problem and also to aim at the recommendation of speech corpora, we work on auto-estimation of the speaking style in a corpus and provide the information as an attribute in the searching system.

To realize the auto-estimation of speaking style, the definition shall be given out first. In 1968, from the aspect of socio-linguistics, Joos indicated that speaking style can be defined according to "casualness" in speeches (Joos, 1968). Then in 1972, Labov mentioned that speaking style changes when the degree of attention that a speaker pays to his or her speech changes (Labov, 1972). Moving to 1990's, Delgado and Freitas indicated that announcers' news reports and teachers' speeches in classrooms, which can be concluded as "professional speech", is also one kind of speech style (Delgado & Freitas, 1991). In Cid and Corugedo's opinion, with or without a speech script shall also be considered in defining speaking style (Cid & Corugedo, 1991). Abe and his colleagues synthesized 3 styles of "artistic novel", "advertisement" and "encyclopedia" by controlling prosodic parameters (Abe, 1994). Based on these studies, Eskenazi proposed that speaking style shall be defined in a data-driven way (Eskenazi, 1993). After reviewing the issues accomplished in the studies that concerned speaking style, Eskenazi proposed 3 compatible scales to capture the nature of speaking style: Intelligibility-Oriented(as "I"), Familiarity(as "F") and soCial strata(as "C"). The scale of Intelligibility-Oriented represents the degree of clarity that the speaker intends his speech to have. It differs from knowing that the listener can catch what the speaker says to a noisy background. Apparently the scale is more about a physical nature. The scale of Familiarity, in literal, means the degree between the speaker and the listener. This scale may differ from identical twins to talking to a foreigner who has little knowledge of the speaker's language and culture. Sometimes, the dialogue context shall also be taken into account. The third scale, Social strata, seems to be more complicated. It stands for the degree of cultivation that the speaker and listener intend to accord their dialogue. It differs from a totally colloquial (lower class) tone to a highly cultivated (upper class) tone. The context of the dialogue and the backgrounds of the speaker and the listener need to be considered in this scale. Comparing to the definition like "casualness" or "artistic novel, advertisement and encyclopedia" mentioned above, we consider that the nature of speaking style can be

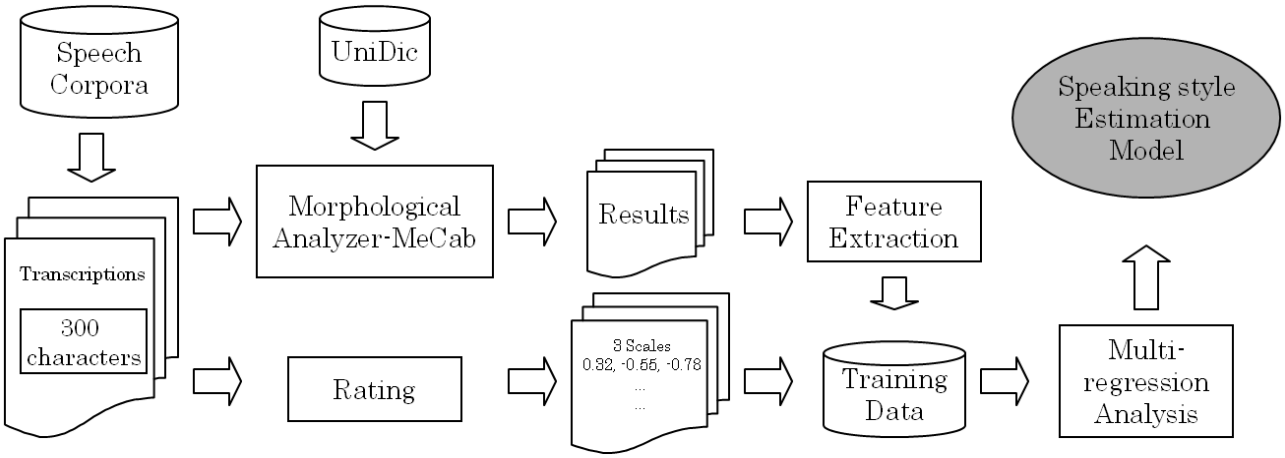

Figure 1: the construction of the speaking style estimation model

specifically described with 3 scales proposed by Eskenazi as an attribute of speech corpora. So, in this paper, we work on auto-estimation in a part of speech corpus to provide the accumulation of speaking style of a whole corpus.

2. Method

As effective factors in auto-estimation of speaking style, both acoustic factors (intonation, pause and etc.) and linguistic factors (morpheme, syntactic structure and etc.) shall be considered. According to Eskenazi, although lots of factors that affect speaking style have been discussed, few studies have been done from the aspect of auto speech processing (Eskenazi, 1993). On the other hand, in the field of natural language processing, style discrimination and author estimation using linguistic features have been actively studied and some satisfactory results have also been achieved(Koiso, 2009)(Koyama, 2008). So in this paper, we attempt to focus on speech transcriptions and use existing methods mentioned above to construct the estimation model of speaking style by referring to the 3 scales of speaking style proposed by Eskenazi.

The process of constructing the estimation model of speaking style is shown in Figure 1.

First, in order to cope with the diversity of speaking styles, we choose several speech samples (including speech transcriptions) from speech corpora randomly. From those speech transcriptions, at the middle part of each speech transcription, we extract about 300 characters as text stimuli to ensure the stability of perceptions of speaking style. Then, to collect the training data for the estimation model, participants are asked to rate for the speaking style perceived in those text stimuli according to Eskenazi' s 3 scales, and the results are to be calculated as the score of speaking style of each text. Morphological analysis using Mecab (Mecab) and UniDic (UniDic) are also to be conducted to extract part of speech, classification of words and morphological patterns, which are useful features for model construction. At last, we construct the estimation model by Multi-regression Analysis and by using the proposing model, we may estimate speaking style of any given speech transcription of any speech samples.

3. Rating

In this section, we introduce the details of rating experiment.

3.1 Participants

22 college students major in information science participate in the rating experiment. None of them is relevant to this study.

3.2 Stimuli

In the rating experiment, we use speech transcriptions in various speech corpora as text stimuli.

3.2.1. Speech Corpora

Considering the cost of the rating experiment and also, to cope with the diversity of speaking style in speech corpora, we randomly choose 10 speech transcriptions each from 6 categories of speech corpora (Shen & Kikuchi, 2012), which are CSJ1, CSJ2, FDC, MAPTASK, AUTO and TRAVEL.

3.2.2. Preprocessing of Text Stimuli

We randomly picked out 10 speech samples from each categories mentioned above and there are totally 60 samples. However, for almost all the samples in speech corpus are longer than about 10 minutes, we extract about 300 characters from the middle part of each speech transcription, which considered enough for perception of speaking style.

Moreover, to avoid the distraction from the contents of each transcription, we replaced every noun (Pronoun is not included) with a "○○" automatically.

3.3 Rating Experiment

The rating experiment is conducted through a CGI on web. All the participants are asked to rate for Eskenazi's 3 scales: Intelligibility-oriented, Familiarity and Social

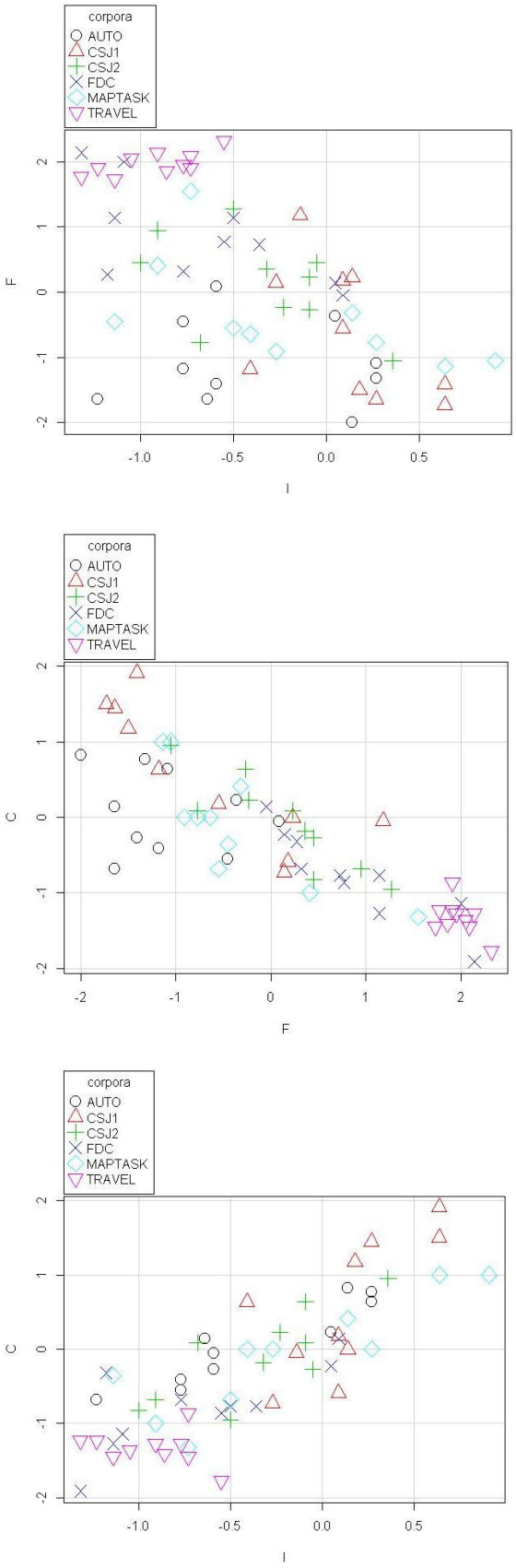

Figure 2: Scatter diagrams of 3 scales
(From top: I-F, F-C, I-C)

strata using a SD method of 7-point after reading each transcription (1 for least intelligible and 7 for very intelligible, 1 for non-familiar and 7 for familiar, 1 for lower strata and 7 for upper strata). The order of texts stimuli for each participant is randomized. However, rereading is allowed and time limit is not set.

3.4 Rating Results

To verify the conformation between the features of 6 categories and the rating results, we observed the distribution of the 60 text stimuli (average of all 22 participants' rating) on the dimension of 3 scales. Figure 2 shows the distributions of each 2 scales. For instance, comparing to the other categories, texts stimuli from TRAVEL distribute at low I and high F. It is because that the speeches in TRAVEL are given by 2 speakers from a same laboratory, which are very familiar with each other and the topic is about making travel plans in a daily life situation. According to Figure 2, it is clear that the results of the rating experiment reflect the features of 6 categories and the diversity of speaking style is ensured as well.

4. Model Construction

In this section, we discuss about model construction using the results of rating experiment and the analysis of texts stimuli.

4.1 Feature Extraction

We use part of speech, classification of words and morphological patterns as features.

4.1.1. Morphological Analysis

We conduct morphological analysis on all the 60 texts stimuli by using Mecab and UniDic.

4.1.2. Part of Speech & Classification of Words

We calculate 10 rates of part of speech (Auxiliary, Verb, Adverb, Pronoun, Adnominal, Conjunction, Particle, Adjective, Interjection and Prefix) and classification of words (function words) in each category.

4.1.3. Morphological Patterns

From speech transcriptions in corpus of spontaneous Japanese, we extracted 43 morphological patterns of linguistic patterns which are considered effective to perceive speaking style (Shen, 2012). In this paper, we cross-checked the 43 patterns with 60 texts stimuli and as a result, 23 matched patterns are used as features to construct estimation model of speaking style (Table 1).

4.2 Estimation Model

With the features mentioned in section 4 as explanatory variables (34 in all), and the average rating results mentioned in section 3 as the objective variable, we construct the multi-regression analysis. There are 3 sub-models representing each scale of the 3 scales of speaking style. We also conduct cross validation (leave-one-out) to verify the reliability of training data. The coefficient of determination (R^2) is shown in Table 2 and details of model are shown in Table 3. In Table 2, we

No.	Morphological Pattern	Sample
1	to[to](par).[iu](v)	toiu
2	to[to](par)shi[suru](v)te[te](par)	toshite
3	yo[yo](par)ne[ne](par)	yone
4	.[keredo](par)mo[mo](par)	keredomo
5	to[to](par)ka[ka](par)	toka
6	.[teru](aux)te[te](par)	ittete
7	tte[tte](par).[iu](v)	tteiu
8	nanka[nanka](par)	nanka
9	ne[ne](par)	ne
10	yo[yo](par)	yo
11	.[zenzen](adv)	zenzen
12	.[yahari](adv)	yappari
13	.[sugoi](adj)	sugoku
14	.[kekkou](adv)	kekkou
15	.[desu](aux) | .[masu](aux)	desu
16	de[de](conj)	de
17	.[konna](adno) | .[sonna](adno) | .[anna](adno)	konna
18	.[chau](aux)	icchau
19	.[toku](aux)	ittoku
20	kocchi[kocchi](pron) | socchi[socchi](pron) | acchi[acchi](pron)	kocchi
21	jya[de](par)	jya
22	politeness	gozaimasu
23	onomatopoeia	hyutto

Table 1: Morphological Patterns
(The notation of morphological patterns: occurrence[lemma](POS))
(“.” is the wildcard and “A|B” means either A or B.)

Features Set	Intelligibility-oriented(I)	Familiarity(F)	soCial strata(C)	Remarks
All features	0.76/0.67	0.93/0.91	0.84/0.79	closed
	0.54/0.37	**0.85**/0.81	**0.73**/0.66	leave-one-out
	0.36/0.13	0.80/0.74	0.62/0.52	leave-one-out(exclusive)
POS only	0.24/0.13	0.32/0.24	0.36/0.31	leave-one-out
	0.14/0.02	-0.08/-0.21	0.23/0.17	leave-one-out(exclusive)
Morph Patterns only	0.36/0.23	0.76/0.67	0.55/0.45	leave-one-out
	0.29/0.14	0.62/0.49	0.44/0.32	leave-one-out(exclusive)

Table 2: Coefficient of determination (R^2 /adjusted R^2)

also show the R^2 by training the estimation model with different feature sets, like using all features, using Part Of Speech only and using morphological patterns only. As a result, the R^2 of using all features are the highest of all (“All features”-“leave-one-out”, I: 0.54, F: 0.85, C: 0.73), which proves the effectiveness of our method. Besides, to observe the possibility of unbalance of features among 6 categories, we also held out the texts stimuli from the same categories in cross validation (leave-one-out). As a result, the R^2 are 0.36 (I), 0.80 (F) and 0.62 (C) (“All

	Intelligibility-oriented(I)		Familiarity(F)		soCial strata(C)	
	explanatory variable	Estimate	explanatory variable	Estimate	explanatory variable	Estimate
1	.[keredo](par)mo[mo](par)**	35.10	.[keredo](par)mo[mo](par),	-25.09	.[keredo](par)mo[mo](par)*	34.06
2	yo[yo](par)ne[ne](par),	19.01	.[desu](aux)| .[masu](aux)***	-19.92	yo[yo](par)***	-21.21
3	tte[tte](par).[iu](v),	-15.16	yo[yo](par)***	19.79	.[chau](aux),	-13.86
4	to[to](par).[iu](v)	12.79	.[chau](aux)*	17.66	.[desu](aux)| .[masu](aux)***	11.34
5	jya[de](par)**	12.33	.[kekkou](adv),	-16.13	adno**	-11.31
6	adno**	-11.30	ne[ne](par)***	14.19	to[to](par)ka[ka](par)*	-9.44
7	.[kekkou](adv)	-11.16	pref,	12.26	.[teru](aux)te[te](par)*	-7.87
8	yo[yo](par),	-9.21	adno*	10.63	aux***	-7.04
9	.[desu](aux)|.[masu](aux)***	7.65	to[to](par)ka[ka](par)*	9.06	ne[ne](par)*	-6.1
10	ne[ne](par)*	-7.44	aux***	6.69	func***	4.89
11	.[teru](aux)te[te](par),	-6.02	func***	-5.79	adj**	-2.84
12	aux***	-4.63	to[to](par).[iu](v)	-5.46	int	-1.51
13	func***	3.67	pron,	4.54	(Intercept)*	-1.78
14	adv,	3.13	adj*	2.69		
15	v,	2.24	(Intercept)***	2.72		
16	par	1.09				
17	(Intercept)***	-3.09				

Table 3: details of the estimation models of speaking style
(Signif. codes: 0 "***", 0.001 "**", 0.01 "*", 0.05 ",", 0.1 " ")
(The notation of morphological patterns: occurrence[lemma](POS))
("." is the wildcard and "A|B" means either A or B.)

features"-"leave-one-out (exclusive)"). Except "I", the results of "F" and "C" are quite satisfactory. In Table 3, the contributive explanatory variables of each sub model (3 scales) are listed in descending order (absolute value). All the sub models are statistically significant at a $p < 0.01$ level.

According to the results above, our proposal of auto-estimation of speaking style is proved effective and by adapting the estimation model on any speech transcription, the speaking style can be estimated in 3 scales.

5. Conclusion

In this study, aiming at recommendation to those users who are interested in utilizing speech corpora, we attempt to estimate the speaking style in speech corpora. We focus on speech transcriptions and use part of speech, classification of words, and morphological patterns, which are indicated to be effective in the field of natural language processing, to construct the estimation model of speaking style by referring to the 3 scales of speaking style proposed by Eskenazi. We construct the estimation model by Multi-regression Analysis. The coefficients of determination of 3 scales are 0.54, 0.85 and 0.73 respectively. The results of Familiarity (F) and soCial strata (C) are satisfactory and indicate the effectiveness of our method. However, the result of Intelligibility-oriented (I) is the lowest in the 3 scales. We consider it might because of lacking of effective features of linguistic factors in speech transcriptions for the Intelligibility-oriented (I) scale. So, as the future work, some other features shall be discussed to improve the model.

6. References

Abe, M., Mizuno, H. (1994). Speaking Style Conversion by Changing Prosodic Parameters and Formant Frequencies. *Proceedings of ICSLP 94*, pp. 1455--1458.

Cid, M., Fernandez Corugedo, S.G. (1991). The Construction of A Corpus of Spoken Spanish: Phonetic and Phonological Parameters. *Proceedings of the ESCA Workshop*, pp. 17-1--17-5.

Delgado, M.R., Freitas, M.J. (1991). Temporal Structures of Speech: Reading News on TV. *Proceedings of the ESCA Workshop*, pp. 19-1--19-5.

ELRA-European Language Resources Association: http://www.elra.info/

Eskenazi, M. (1993). Trends in Speaking Style Research. Keynote speech, *Proceedings of Eurospeech'93*.

Joos, M. (1968). The Isolation of Styles. In FISHMAN, J.A. (ED) Readings in *the Sociology of Language*, The Hague: Mouton, pp. 185--191.

Jorden, E., Noda, M. (1987). *Japanese the Spoken Language*. New Haven & London: Yale University Press.

Koiso, H., Ogiso, T., Miyauchi, S. (2009). Title omitted (in Japanese). *Proceedings of the 15th Annual*

conference of the association for Natural Language Processing, pp. 594--597.

Koyama, T., Takeuchi, K. (2008). An Evaluation of Document Set Similarity Based on Morpheme Occurrence Patterns. *Information Processing Society of Japan, Technical Report*, NL-188, pp. 51--55.

Labov, W. (1972). The Isolation of Contextual Styles. *Sociolinguistic Patterns*, Oxford, pp. 70--109.

LDC- Linguistic Data Consortium: http://www.ldc.upenn.edu/

Mecab: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

NII-SRC-Speech Resources consortium: http://research.nii.ac.jp/src/

Shen, R., Kikuchi, H. (2011). Construction of the Speech Corpus Retrieval System: Corpus Search & Catalog-Search. *Proceedings of O-COCOSDA 2011*, pp. 76--80.

Shen, R., Kikuchi, H. (2012). Ratings of Speaking Style in Speech Corpora - Focus on Transcriptions. *Proceedings of O-COCOSDA 2012*, pp. 274--278.

Shen, R., Kikuchi, H., Ohta, K., Mitamura, T. (2012). Towards the Text-level Characterization Based on Speech Generation. *Journal of Information Processing Society of Japan*, 53-4, pp. 1269--1276.

UniDic: http://www.tokuteicorpus.jp/dist/

Yamakawa, K., Kikuchi, H., Matsui, T., Itahashi, S. (2009). Utilization of Acoustical Feature in Visualization of Multiple Speech Corpora. *Proceedings of Oriental COCOSDA 2009*.