早稲田大学審査学位論文

博士（人間科学）

概要書

Automatic Estimation of Speaking Style in Speech Corpora

２０１４年７月

早稲田大学大学院　人間科学研究科

沈　睿

SHEN, Rui

研究指導教員：　菊池　英明　教授

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. In Japan, we cooperated with 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 (K. Yamakawa 2009) (R. Shen 2011). However, besides attributes like "purpose" or "speakers", "speaking style" shall also be considered useful information. According to Jorden (E. 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. Eskenazi proposed that speaking style shall be defined in a data-driven way (M. 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.

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 (M. 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(H. Koiso 2009) (T. 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.

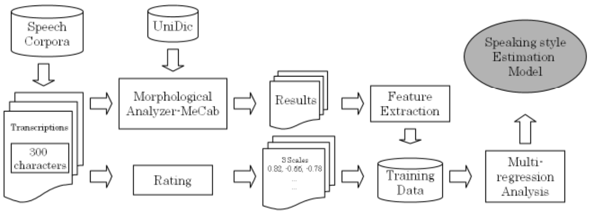
Speech Corpora
UniDic
Speaking style Estimation Model
Transcriptions
300 characters
Morphological Analyzer-MeCab
Results
Feature Extraction
Rating
3 Scales 0.82, -0.66, -0.78 … …
Training Data
Multi-regression Analysis

Figure 1: the construction of the speaking style estimation model

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 and 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.

Rating

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

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

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

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 (R. Shen 2012), which are CSJ1, CSJ2, FDC, MAPTASK, AUTO and TRAVEL.

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.

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 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.

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.

Model

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

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

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

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.

From speech transcriptions in corpus of spontaneous Japanese, we extracted 43 morphological patterns of linguistic patterns which are considered effective to perceive speaking style (R. 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.

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. W lso conduct cross validation (leave-one-out) to verify the reliability of training data. The coefficients of determination (R2) are I: 0.54, F: 0.85 and C: 0.73), which proves the effectiveness of our method. 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.

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 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.

References

M. Eskenazi, “Trends in Speaking style Research.” Keynote speech, Proc. of Eurospeech’93, Berlin, 1993.

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

H. Koiso, T. Ogiso and S. Miyauchi, “A Corpus-based Stylistic Comparison on Various Genres: Focusing on Short-Unit Word.” (in Japanese, the title translated by the author), Proc. of the 15th Annual conference of the association for Natural Language Processing, Vol.15, pp.594-597, 2009.

T. Koyama and K. Takeuchi, “An Evaluation of Document Set Similarity Based on Morpheme occurrence Patterns.” IPSJ SIG Technical Report, NL-188 (8), pp.51-56, 2008.

R. Shen and K. Hideaki, “Construction of the Speech Corpus Retrieval System: Corpus Search & Catalog-Search.” Proc. of Oriental-COCOSDA 2011, pp.76-80, 2011.

R. Shen and H. Kikuchi, “Ratings of Speaking Style in Speech Corpora - Focus on Transcriptions.” Proc. of Oriental-COCOSDA 2012, pp.274-278, 2012.

R. Shen, H. Kikuchi, K. OHTA and T. MITAMURA, “Towards the Text-level Characterization Based on Speech Generation.” Journal of Information Society of Japan, Vol.53, No.4, pp.1269-1276, 2012.

K. Yamakawa, H. Kikuchi, T. Matsui and S. Itahashi, “Utilization of Acoustical Feature in Visualization of Multiple Speech Corpora.” Proc. of Oriental COCOSDA 2009, Beijing, China, pp. 147-151, 2009.