論文 / 著書情報
Article / Book Information

Title	Analysis on Individual Differences in Automatic Transcription of Spontaneous Presentations
Authors	Takahiro Shinozaki, Sadaoki Furui
Citation	IEEE ICASSP 2002, Vol. 1, No. SP-P11.07, pp. 729-732
Pub. date	2002, 5
Copyright	(c) 2002 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
URL	http://www.ieee.org/index.html
DOI	http://dx.doi.org/10.1109/ICASSP.2002.5743821
Note	This file is author (final) version.

ANALYSIS ON INDIVIDUAL DIFFERENCES IN AUTOMATIC TRANSCRIPTION OF SPONTANEOUS PRESENTATIONS

Takahiro Shinozaki and Sadaoki Furui

Tokyo Institute of Technology
Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan.
{staka, furui}@furui.cs.titech.ac.jp

ABSTRACT

This paper reports an analysis of individual differences in spontaneous presentation speech recognition performances. Ten minutes from each presentation given by 50 male speakers, for a total of 500 minutes, has been automatically recognized for the analysis. Correlation and regression analyses were applied to the word recognition accuracy and various speaker attributes. A restricted set of the speaker attributes comprising the speaking rate, the out of vocabulary rate and the repair rate was found to be most significant to yield individual differences in the word accuracy. Unsupervised MLLR speaker adaptation worked well for improving the word accuracy but did not change the structure of the individual differences. Approximately half of the variance in the word accuracy was explained by a regression model using the limited set of three attributes.

1. INTRODUCTION

To promote better understanding and to build technology for spontaneous speech recognition, the Science and Technology Agency Priority Program (Organized Research Combination System) entitled ”Spontaneous Speech: Corpus and Processing Technology” started in 1999 under the supervision of Furui [1]. A large-scale spontaneous speech corpus named "Corpus of Spontaneous Japanese (CSJ)" is under construction by the project. Previous study showed that acoustic and language models made using the CSJ were significantly superior to conventional read-speech-based models when applied to spontaneous speech recognition [2]. However, the recognition accuracy is still rather low, and there might be many factors that affect recognition performance acoustically as well as linguistically.

It is presumable that variation of speaking style is larger in spontaneous speech than in read speech according to the degree of speaker's freedom. And so does the word accuracy. Knowing the structure of speaking style differences among individuals and the influence on word accuracy it exerts is very important to promote spontaneous speech recognition systems. This paper reveals the structure of individual differences in the word accuracy based on recognition results in presentation speech uttered by 50 male speakers processed by a state-of-the-art recognition system.

Section 2 describes the task and experimental set up. Experimental results and analyses are presented in Section 3. Finally some conclusions are given in Section 4.

2. SPEECH RECOGNITION TASK AND EXPERIMENTAL SET UP

2.1. Recognition task

For the analysis of speaker variation, monologue presentation speech uttered by 50 different male speakers is used as a test set. Speakers in the test set have no overlap with those in the training set. The first 10 minutes of each presentation are used for analysis. Table 1 shows the detail of the test set.

2.2. Speaker attributes

We give consideration to seven kinds of speaker attributes in the analysis. They are word accuracy (Acc), averaged acoustic frame likelihood (AL), speaking rate (SR), word perplexity (PP), out of vocabulary rate (OR), filled pause rate (FR) and repair rate (RR).

The speaking rate which we define as the number of phonemes per second and the averaged acoustic frame likelihood are calculated using the result of forced alignment of the reference tri-phone label after removing pause periods. Word perplexity is calculated using tri-grams, in which prediction of out of vocabulary words is not included. The filled pause rate and the repair rate are the percentage of filled pauses and repairs in total words, respectively. Tag information included in CSJ transcription is used to determine whether a word is a filled pause/repair or not. In CSJ, repairs are defined only for word fragments, and a whole word which is rephrased is not marked as a repair. The calculations of word accuracy, out of vocabulary rate and word perplexity are based on the reference sentence after excluding repairs.

2.3. Experimental conditions

Speech signals are digitized with 16kHz sampling and 16bit quantization. Feature vectors have 25 elements consisting of 12 MFCC, their delta and the delta log energy. The CMS (cepstral mean subtraction) is applied to each utterance. HTK v2.2 is used for acoustic modeling and adaptation. Language models are made by using the CMU SLM Tool Kit v2.05. Morphemes (which will be called ”words” hereafter in this paper) are used as units for statistical

Table 1. Test set

Conference	No. presentations
Jap. Soc. AI	32
Acoust. Soc. Jap.	12
Others	6

language modeling. The Julius v3.1 decoder [3] is used for speech recognition.

2.4. Language and acoustic modeling

A part of the CSJ so far completed, having approximately 1.5M words, is used as a training set. The training set consists of 610 presentations; 274 academic conference presentations and 336 simulated presentations.

The language model used in the recognition consists of bi-grams and reverse tri-grams with backing-off. It is made using the whole training set. The vocabulary size is 30k. We treated filled pauses as words in modeling. Repairs are deleted from training text and are not modeled. This is because modeling repairs effectively by N-gram is difficult due to the large amount of variations and few occurrences of each fragment.

A speaker independent (SI) acoustic model is made using 338 presentations uttered by male speakers (approximately 59 hours). It is a tied-state tri-phone HMM having 2k states and 16 Gaussian mixtures in each state. Each tri-phone HMM has three states with the left-to-right structure.

In addition, we incorporate a batch-type unsupervised speaker adaptation to see the effect on the individual differences. We apply the MLLR method in which a regression class tree having 64 leaves is made using a centroid-splitting algorithm. We denote the resulting set of speaker adaptive HMMs for 50 speakers as SA HMMs.

The language model weights and the insertion penalties are chosen to maximize the recognition accuracy of the test set for each combination of the SI/SA acoustic model and language model but kept constant for all test speakers.

3. ANALYSIS OF THE STRUCTURE OF INDIVIDUAL DIFFERENCES

Table 2 shows the mean and standard deviation over the 50 speakers for the word accuracy and other six kinds of the speaker attributes. The calculation of the speaking rate is based on the SI HMM. The mean word accuracy of the 50 speakers is 64.2% and 68.6% for the SI and SA conditions respectively. The standard deviation is 7.4% for the SI and 7.5% for the SA condition. As shown by the standard deviation, recognition accuracy largely varies from speaker to speaker. We discuss correlation analysis in 3.1 and regression analysis in 3.2.

3.1. Correlation analysis

Table 3 shows the correlation matrix of speaker attributes. In the table, the lower triangular matrix shows the correlation coefficients and the upper triangular matrix shows the observed significance levels (p-values). The correlation coefficients written in bold face indicate significant values at 5% significance level (p-value < 0.05).

3.1.1. Correlation between acoustic likelihood and speaking rate

The correlation coefficient between acoustic likelihood and speaking rate is -0.59 for the SI acoustic model. Figure 1 shows the relationship between the speaking rate and the averaged frame likelihood. There is a tendency that the higher the speaking rate is, the lower the acoustic likelihood becomes. On the other hand, even very slow speaking rate does not cause decrease of the acoustic likelihood. The Akaike Information Criterion (AIC) [4] also indicates that the first order regression model is better than the second order model for regressing the acoustic likelihood on the speaking rate. This indicate that there is a linear relationship between the speaking rate and the acoustic likelihood averaged over presentations. A stronger articulation effect in faster speakers is probably a cause of the decrease of likelihood.

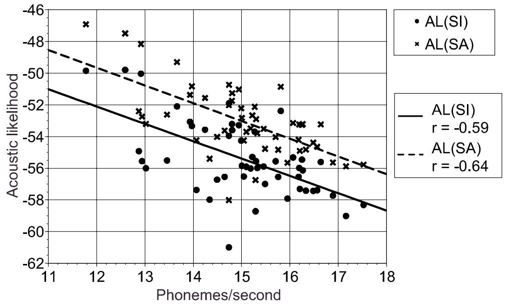

Fig. 1. Speaking rate vs. acoustic likelihood.

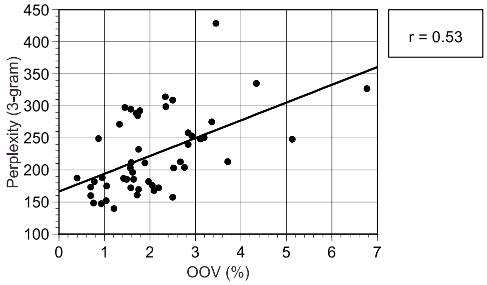

Fig. 2. OOV vs. word perplexity.

The unsupervised adaptation increases the acoustic likelihood but keeps the relationship between the speaking rate and the acoustic likelihood with a slight increase in the correlation coefficient.

3.1.2. Correlation between word perplexity and several linguistic attributes

There exists significant correlation between the word perplexity and the out of vocabulary rate with the correlation coefficient of 0.53. Figure 2 shows the relationship between the word perplexity and the out of vocabulary rate. There is a tendency that presentations having a higher out of vocabulary rate show a higher perplexity.

The correlation coefficient of the filled pause frequency and the perplexity is -0.19 indicating that they are almost uncorrelated. The repair frequency and the perplexity has a correlation coefficient of 0.11. Since the perplexity was calculated after removing repairs, this result shows that the linguistic difficulty excluding repairs has almost no correlation with the repair rate.

3.1.3. Correlation between word accuracy and several attributes

The correlation coefficient between the word accuracy (SI) and the speaking rate is -0.47. Figure 3 shows the relationship between the word accuracy and the speaking rate. The relationship seems monotonic and even very slow speaking rate does not decrease the accuracy, which is similar to the result for the acoustic likelihood shown in Figure 1. The AIC also indicates that the first

Table 2. Mean and standard deviation for each attribute

	Acc(SI)	Acc(SA)	AL(SI)	AL(SA)	SR	PP	OR	FR	RR
Mean	64.2	68.6	-55.4	-53.1	15.0	224	2.09	8.59	1.56
Standard deviation	7.4	7.5	2.3	2.2	1.2	61	1.18	3.67	0.72

Table 3. Correlation coefficient matrix; the lower triangular matrix shows the correlation coefficients and the upper triangular matrix shows the *p*-value, that is, the significance level. Bold face indicates a significant value with the significant level of 5%

	Acc(SI)	Acc(SA)	AL(SI)	AL(SA)	SR	PP	OR	FR	RR
Acc(SI)		–	5.4%	–	0.1%	0.5%	0.0%	0.6%	2.2%
Acc(SA)	–		–	2.4%	0.0%	1.6%	0.0%	0.6%	2.4%
AL(SI)	0.27	–		–	0.0%	65.1%	12.5%	6.9%	46.9%
AL(SA)	–	**0.32**	–		0.0%	52.3%	8.6%	7.0%	34.7%
SR	**-0.47**	**-0.49**	**-0.59**	**-0.64**		65.1%	1.2%	0.0%	34.0%
PP	**-0.39**	**-0.34**	-0.07	-0.09	0.07		0.0%	18.0%	44.8%
OR	**-0.54**	**-0.51**	-0.22	-0.25	**0.35**	**0.53**		0.3%	67.9%
FR	**0.38**	**0.38**	0.26	0.26	**-0.51**	-0.19	**-0.41**		32.9%
RR	**-0.32**	**-0.32**	-0.10	-0.14	0.14	0.11	-0.06	0.14	

order model is superior to the second order model for regressing the word accuracy on the speaking rate.

Correlation between the word accuracy and the acoustic likelihood is not statistically significant, when the SI acoustic model is used. Their partial correlation coefficient adjusted for the speaking rate is -0.005. Partial correlation coefficient between the word accuracy and the speaking rate adjusted for the acoustic likelihood is -0.40, which is significant at a 1% significance level, and partial correlation coefficient between the acoustic likelihood and the speaking rate adjusted for the word accuracy is -0.54, which is significant at a 1% significance level. This means that the correlation between the word accuracy and the acoustic likelihood is spurious. In other words, a fast speaking rate decreases the word accuracy and the acoustic likelihood independently. Similar results are obtained for the SA conditions.

The correlation coefficient between the word accuracy and the repair frequency is -0.32. Figure 4 shows the scattergram of the word accuracy and the repair rate when the SI acoustic model is used.

There is a weak positive correlation of 0.38 between the word accuracy and the filled pause frequency, but this is also a spurious correlation, since partial correlation coefficient adjusted for the speaking rate is 0.18.

Figure 5 shows the scattergram of the word accuracy (SI) and the out of vocabulary rate. The correlation coefficient between the word accuracy and the out of vocabulary rate is -0.54.

There is a weak negative correlation of -0.39 between the word accuracy (SI) and the perplexity, but this is also spurious; the partial correlation between the word accuracy and the perplexity adjusted for the out of vocabulary rate is -0.14.

3.2. Regression analysis

The following equations (1) and (2) show linear regression models of the word accuracy with the six presentation attributes when the SI and SA acoustic model are respectively used for speech recognition.

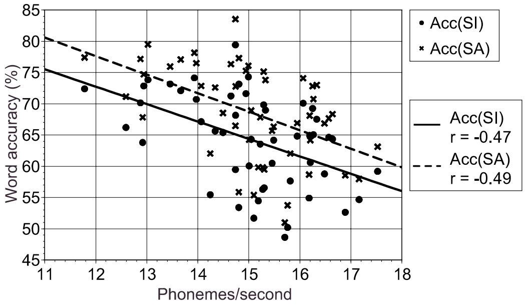

Fig. 3. Speaking rate vs. word accuracy.

$$\begin{aligned} Acc_{SI} &= -0.061AL_{SI} - 1.4SR_{SI} - 0.014PP \\ &\quad -2.3OR + 0.28FR - 3.3RR + 92 \qquad (1) \end{aligned}$$

$$\begin{aligned} Acc_{SA} &= -0.061AL_{SA} - 1.6SR_{SI} - 0.010PP \\ &\quad -2.1OR + 0.30FR - 3.3RR + 98 \qquad (2) \end{aligned}$$

In the equation (1), regression coefficient for the repair rate is -3.3 and the coefficient for the out of vocabulary rate is -2.3. This means that 1% increase of the repair rate or the out of vocabulary rate respectively corresponds to 3.3% or 2.3% decrease of the word accuracy. This is probably because single recognition error caused by a repair or an out of vocabulary word triggers secondary errors due to the linguistic constraints. Regression coefficients before and after speaker adaptation are almost the same excepting the constant term. The coefficient of determination for the multiple linear regression (1) is 0.50 and that for (2) is 0.47, both are significant at 1% level. This means that about a half of the variance of the word accuracy is explained by the model.

Table 4 shows standardized representation of the regression analysis with the equations (1) and (2), in which the variables are

Table 4. Standardized regression analysis results, showing standardized regression coefficient (Coeff), *p*-value and 95% confidence interval (95% CI).

	Coeff(SI)	P	95% CI		Coeff(SA)	P	95% CI
AL(SI)	-0.02	0.885	(-0.29, 0.25)	AL(SA)	-0.02	0.904	(-0.31, 0.28)
SR(SI)	-0.23	0.149	(-0.55, 0.09)	SR(SI)	-0.26	0.135	(-0.60, 0.08)
PP	-0.12	0.374	(-0.38, 0.15)	PP	-0.08	0.549	(-0.36, 0.19)
OR	-0.36	0.015	(-0.65,-0.07)	OR	-0.33	0.028	(-0.63,-0.04)
FR	0.14	0.305	(-0.13, 0.41)	FR	0.15	0.301	(-0.14, 0.43)
RR	-0.32	0.008	(-0.55,-0.09)	RR	-0.32	0.010	(-0.55,-0.08)

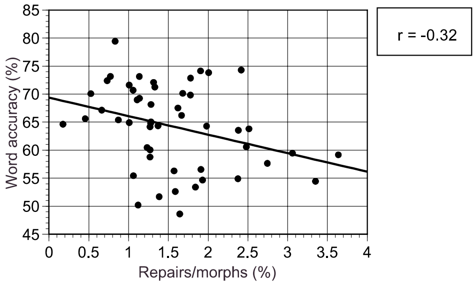

Fig. 4. Repair frequency vs. word accuracy (SI).

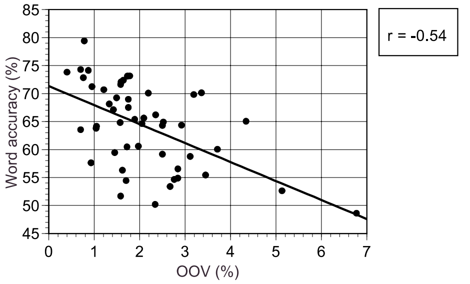

Fig. 5. OOV rate vs. word accuracy (SI).

standardized before the analysis in order to show the effects of explaining variables on the word accuracy. The table shows the standardized regression coefficient, the *p*-value and the 95% confidence interval. The standardized regression coefficients of the acoustic likelihood, the perplexity and the filled pause rate are relatively small for both the SI and SA regression models. Although most of these variables have statistically significant correlation with the word accuracy, these correlation are spurious as indicated in Subsection 3.1.

3.3. Discussion

As a supplementary experiment, we employed a backward elimination procedure to identify relatively important predictors of the word accuracy. A backward elimination process started with all of the six predictors in the model, and the model was refitted to the data after removing a variable with the largest *p*-value. The refitting process was iterated removing the least significant variable in the model until all remaining variables had *p*-values smaller than 0.05. The important predictors identified were the speaking rate, the out of vocabulary rate and the repair rate, which correspond to the attributes showing relatively large coefficient in Table 4. Coefficients of determination of the regression models on these three attributes are 0.48 and 0.46 for speaker independent and adaptive cases, that are almost the same as that of the models on all attributes. It can be concluded that main factors of individual differences of the word accuracy are the speaking rate, the out of vocabulary rate and the repair rate.

4. CONCLUSION

In this paper, we have investigated the individual differences in spontaneous presentation speech recognition. It was shown that the speaking rate, the out of vocabulary rate and the repair rate have relatively large effects on the individual differences of the word accuracy among a set of presentation/speaker attributes. We have found that the averaged acoustic likelihood of reference phoneme sequences and the test set perplexity are relatively minor factors of individual differences in the word accuracy for the 50 male speakers in the test set.

Unsupervised MLLR speaker adaptation works well for improving the word accuracy but do not change the structure of the individual differences including the effects of the speaking rate. A special method for addressing speaking rate is crucial.

Approximately half of the variance of the word accuracy is explained by the regression model on the set of six explaining variables. The regression model on the three most important attributes also displays a similar prediction power.

Our future research includes investigation of efficient methods for reducing the effects of the major attributes on the recognition accuracy.

5. REFERENCES

[1] S. Furui, et al., “Toward the realization of spontaneous speech recognition,” Proc. ICSLP, China, Vol. 3, pp.518-521, 2000.

[2] T. Shinozaki, C. Hori, and S. Furui, “Towards automatic transcription of spontaneous presentations,” Proc. EUROSPEECH, Denmark, Vol. 1, pp.491-494, 2001.

[3] A. Lee, et al., “An efficient two-pass search algorithm using word trellis index,” Proc. ICSLP, Australia, pp.1831-1834, 1998.

[4] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” Proc. ISIT, (B. N. Petrov and F. Csaki eds.) Akademiai Kiado, Budapest, pp.267-281, 1973.