論文 / 著書情報
Article / Book Information

Title(English)	Why is Automatic Recognition of Spontaneous Speech So Difficult?
Authors(English)	Sadaoki Furui, Masanobu Nakamura, Koji Iwano
Citation(English)	International Symposium on Large-Scale Knowledge Resources (LKR 2006), Vol. , No. , pp. 83-90
発行日 / Pub. date	2006, 3

Proc. International Symposium on Large-scale Knowledge Resources (LKR2006), 2006.3

Why is Automatic Recognition of Spontaneous Speech So Difficult?

Sadaoki Furui, Masanobu Nakamura, and Koji Iwano

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

Abstract

Although speech derived from reading texts, and similar types of speech, e.g. that from reading newspapers or that from news broadcast, can be recognized with high accuracy, recognition accuracy drastically decreases for spontaneous speech. This is due to the fact that spontaneous speech and read speech are significantly different acoustically as well as linguistically. This paper reports analysis and recognition of spontaneous speech using a large-scale spontaneous speech database "Corpus of Spontaneous Japanese (CSJ)". Recognition results in this experiment show that recognition accuracy significantly increases as a function of the size of acoustic as well as language model training data and the improvement levels off at approximately 7M words of training data. This means that a very large corpus is needed to encompass the huge linguistic and acoustic variations which occur in spontaneous speech. Spectral analysis using various styles of utterances in the CSJ shows that the spectral distribution/difference of phonemes is significantly reduced in spontaneous speech compared to read speech. Experimental results also show that there is a strong correlation between mean spectral distance between phonemes and phoneme recognition accuracy. This indicates that spectral reduction is one major reason for the decrease of recognition accuracy of spontaneous speech. Comparative analysis of statistical language models for written language, including newspaper articles, and spontaneous speech shows that there is a significant difference between written language and spontaneous speech in terms of observation frequency of each part-of-speech and perplexity.

1. Introduction

State-of-the-art speech recognition technology can achieve high recognition accuracies for read texts or limited domain spoken interactions. However, the accuracy is still rather poor for spontaneous speech, which is not as well structured acoustically and linguistically as read speech [1, 2]. Spontaneous speech includes filled pauses, repairs, hesitations, repetitions, partial words, and disfluencies. It is quite interesting to note that, although speech is almost always spontaneous, until recently speech recognition research has focused primarily on recognizing read speech. Spontaneous speech recognition as a specific research field has only recently emerged in the past 10 years within the wider field of automatic speech recognition (e.g. [3, 4, 5, 6, 7]). Effectively broadening the application of speech recognition depends crucially on raising recognition performance for spontaneous speech.

In order to increase recognition performance for spontaneous speech, it is necessary to build acoustic and language models specific to spontaneous speech. Our knowledge of the structure of spontaneous speech is currently insufficient to achieve the necessary breakthroughs. Although spontaneous speech phenomena are quite common in human communication and may increase in human machine discourse as people become more comfortable conversing with machines, analysis and modeling of spontaneous speech are only in the initial stages. It is widely well known that spectral distribution of continuously spoken vowels or syllables is much smaller than that of isolated spoken vowels or syllables, which is sometimes called spectral reduction. Similar reduction has also been observed for spontaneous speech in comparison with read speech (e.g. [8, 9]). However, as of yet no research has been conducted using a large spontaneous database nor has there been research studying the relationship between the spectral reduction and spontaneous speech recognition performance. As for language modeling, no research has been conducted comparing written language and spontaneous speech by using a large-scale corpus.

The next section in this paper introduces our large-scale spontaneous speech corpus, the Corpus of Spontaneous Japanese, and the following sections report results of experiments using the spontaneous speech corpus. The experiments were conducted to evaluate the effectiveness of the spontaneous speech corpus on speech recognition, to investigate spectral reduction using cepstral features that are widely used in speech recognition, to analyze the differences of distance between each pair of phonemes in spontaneous speech and that in read speech, to investigate the relationship between the phoneme distance and phoneme recognition performance in various speaking styles, and to compare language models for written text and spontaneous speech using a newspaper corpus and news commentary text in addition to the various text included in the spontaneous speech corpus.

2. Corpus of Spontaneous Japanese (CSJ)

A 5-year Science and Technology Agency Priority Program entitled "Spontaneous Speech: Corpus and Processing Technology" was conducted in Japan from 1999 to 2004 [1] to build a large-scale spontaneous speech corpus, Corpus of Spontaneous Japanese (CSJ), consisting of roughly 7M words (morphemes) with a total speech length of 650 hours [10, 11]. Mainly recorded in the CSJ are monologues such as academic presentations (AP) and extemporaneous presentations (EP) as shown in Table 1. AP contains live recordings of academic presentations in nine different academic societies covering the fields of engineering, social science and humanities. EP is studio recordings of paid layman speakers' speech on everyday topics like "the most delightful memory of my life" presented in front of a small audience and in a relatively relaxed atmosphere. Therefore, the speaking style in EP is more informal than in AP. Presentations reading text have been excluded from AP and EP. The EP recordings provide a more balanced representation of age and gender than the AP. The CSJ also includes a smaller

Table 1: Contents of the CSJ.

Type of speech	# speakers	# files	Monologue/ Dialogue	Spontaneous/ Read	Hours
Academic presentations (AP)	838	1006	Monolog	Spont.	299.5
Extemporaneous presentations (EP)	580	1715	Monolog	Spont.	327.5
Interview on AP	*(10)	10	Dialog	Spont.	2.1
Interview on EP	*(16)	16	Dialog	Spont.	3.4
Task oriented dialogue	*(16)	16	Dialog	Spont.	3.1
Free dialogue	*(16)	16	Dialog	Spont.	3.6
Reading text	*(244)	491	Dialog	Read	14.1
Reading transcriptions	*(16)	16	Monolog	Read	5.5
*Counted as the speakers of AP or EP				Total hours	658.8

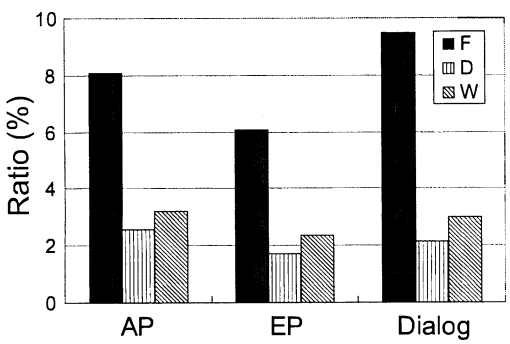

Figure 1: Ratios of filled pauses (F), word fragments (D) and reduced articulation or mispronunciation (W) in AP, EP and dialogue.

database of dialogue speech for the purpose of comparison with monologue speech. The dialogue speech is composed of an interview, a task oriented dialogue, and a free dialogue. The "reading text" in the table indicates the speech reading novels including dialogues, and the "reading transcriptions" indicates the speech reading transcriptions of APs or EPs by the same speaker. The recordings were manually given orthographic and phonetic transcription. Spontaneous speech-specific phenomena, such as filled pauses, word fragments, reduced articulation or mispronunciation, and non-speech events like laughter and coughing were also carefully tagged.

One-tenth of the utterances, hereafter referred to as the Core, were tagged manually and used for training a morphological analysis and part-of-speech (POS) tagging program [12] for automatically analyzing all of the 650-hour utterances. The Core consists of 70 APs, 107 EPs, 18 dialogues and 6 read speech files (speakers). They were also tagged with paralinguistic/intonation information, dependency-structure, discourse structure, and summarization. For intonation labeling of spontaneous speech, the traditional J_ToBI method [13] was extended to X_JToBI [14], in which inventories of tonal events as well as break indices were considerably enriched.

Figure 1 shows mean values of the ratio of disfluencies, specifically filled pauses (F), word fragments (D), and reduced articulation or mispronunciation (W), to the total number of words included in AP, EP and dialogues (interviews, task oriented dialogues and free dialogues), respectively. These results show that approximately one-tenth of the words are disfluencies in the spontaneous speech in the CSJ, and there is no significant difference among the overall ratios of disfluencies in terms of AP, EP or dialogues. It is also observed that the ratio of F is significantly higher than that of D and W.

3. Progress made and difficulties encountered in spontaneous speech recognition

3.1. Test sets for technology evaluation

In order to evaluate spontaneous speech recognition technology, three test sets of presentations were constructed from the CSJ to properly represent the whole corpus with respect to various factors of spontaneous speech [15]. An analysis by Shinozaki et al. [16] concluded that speaking rate (SR), out-of-vocabulary (OOV) rate (OR) and repair rate (RR) were three major speaker attributes highly correlated with accuracy. Other factors primarily depended on one or more of the three factors. For example, word perplexity (PP) was also highly correlated with accuracy, but if its correlation with the OR was removed, they found actually that the correlation between PP and accuracy was significantly reduced. However, OR is intrinsically dependent on vocabulary and is thus variable when the lexicon is modified. On the other hand, the difference of PPs among speech data is generally more stable, even when the language model is revised. Therefore, we decided to take into account PP instead of OR, in combination with SR and RR, in the test-set selection.

Since the speaking styles and vocabularies of AP and EP are significantly different, we set up respective test sets. In addition, considering the fact that most of the AP presentations were given by male speakers, we set up two sets for the academic category: a male-only set and a gender-balanced set. Thus, we constructed three test sets, each of which consisted of 10 speakers: male speakers AP, gender-balanced AP, and gender-balanced EP. The remaining AP as well as EP presentations, excluding those having overlap with the test sets in terms of speakers, were set up as training data (510 hours, 6.84 M words). The utterances were digitized by 16 kHz and converted into a sequence of feature vectors consisting of MFCC (Mel-frequency cepstrum coefficients), ΔMFCC and Δlog-energy features, using a 25 ms-length window shifted every 10 ms. Benchmark re-

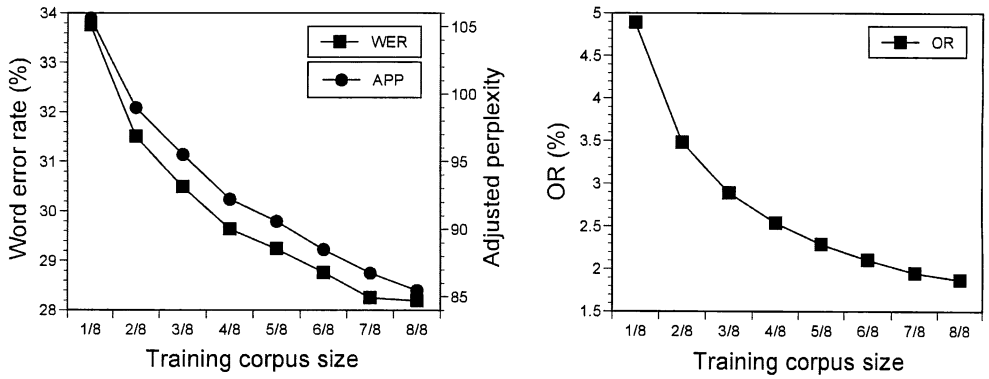

Figure 2: Word error rate (WER), adjusted test-set perplexity (APP) and out-of-vocabulary (OOV) rate (OR) as a function of the size of language model training data.

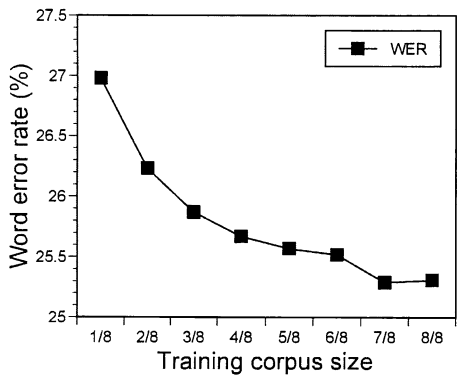

Figure 3: WER as a function of the size of acoustic model training data.

sults of speech recognition using these three test sets have also been presented in our previous paper [15].

3.2. Effectiveness of corpora

By constructing acoustic and language models using the CSJ, recognition errors for spontaneous presentation were reduced to roughly half compared to models constructed using read speech and written text [1, 3]. Increasing the size of training data for acoustic and language models decreased the recognition error rate (WER: word error rate) as shown in Figures 2 and 3 [17]. They show the results averaged over the three test sets. Figure 2 indicates WER, adjusted test-set perplexity (APP) [18] and OOV rate (OR), as a function of the size of language model training data with the condition that the acoustic model is constructed using the whole training data set (510 hours). The adjusted perplexity was used, since it normalizes the effect of the reduction of OOV rate on the perplexity according to the increase of training data size. On the other hand, Figure 3 shows WER as a function of the size of the acoustic model training data, when the language model is made using the whole training data set (6.84M words).

By increasing the language model training data size from 1/8 (0.86M words) to 8/8 (6.84M words), the WER, the perplexity and the OOV are relatively reduced by 17%, 19%, and 62%, respectively. By increasing the acoustic model training data from 1/8 (68 hours) to 8/8 (510 hours), the WER is reduced by 6.3%. The best WER of 25.3%, obtained by using the whole training data set for both acoustic and language modeling, shown at the extreme right condition in Figure 3, is 2.9% lower in the absolute value than that shown in Figure 2. This is because the former experiment of Figure 3 combined $\Delta\Delta$MFCC and $\Delta\Delta$log-energy with the three features of MFCC, ΔMFCC and Δlog-energy which were used in the experiment of Figure 2. All these results show that WER is significantly reduced by an increase of the size of training data and almost saturated by using the whole data set. This strongly confirms that the size of the CSJ is meaningful in modeling spontaneous presentation speech using the standard model training strategies.

4. Spectral space reduction in spontaneous speech and its effects on speech recognition performances

4.1. Spectral analysis of spontaneous speech

Results of recognition experiments using the spontaneous presentations in the CSJ clearly show that spontaneous speech and read speech are acoustically different. In order to clarify the acoustical differences, spectral characteristics of spontaneous speech have been analyzed in comparison with that of read speech [19, 20]. Utterances in various speaking styles (speaking types) in the CSJ, such as AP, EP, utterances reading the transcription of AP (read transcription speech), and dialogues, were used in the analysis. The dialogue utterances consisted of interviews concerning AP, interviews concerning EP, task dialogues, and free dialogues. In order to remove the effect of individual differences, utterances in different styles by the same five male and five female speakers were compared. Since not only the speakers but also the text were identical for the reading of the transcribed speech and the original AP utterances, very precise comparative analysis could be performed. The total number of phoneme samples used in this experiment for each speaker and each speaking style varied between 4,837 and 29,862. Each presentation had a duration of 10 minutes in average.

These utterances were segmented by silences with dura-

Table 2: Japanese phonemes.

Vowel	/a,i,u,e,o,a:,i:,u:,e:,o:/
Consonant	/w,y,r,p,t,k,b,d,g,j,ts,ch, z,s,sh,h,f,N,N:,m,n/

tions of 400 ms or longer. If the length of the segmented unit was shorter than 1 sec, it was merged with the succeeding unit. The segmented utterances are hereafter called "utterance units".

The whole set of 31 Japanese phonemes, consisting of 10 vowels and 21 consonants, are listed in Table 2. The mean and variance of MFCC vectors for each phoneme in various speaking styles were calculated to analyze the spectral characteristics of spontaneous speech as follows.

1. 39-dimensional feature vectors, consisting of 12-dimensional MFCC, log-energy, and their first and second derivatives, were extracted from utterances using a 25 ms-length window shifted every 10 ms. The CMS (cepstral mean subtraction) was applied to each utterance unit.
2. A mono-phone HMM with a single Gaussian mixture was trained using utterances of every combination of phonemes, speakers, and utterance styles. Every HMM had a left-to-right topology with three self-loops.
3. The mean and variance vectors of the 12-dimensional MFCC at the second state of the HMM were extracted for each phoneme and used for the analysis.

4.2. Projection into the PCA space

The distribution of mean MFCC vectors of all the vowels and consonants for the dialogue and read speech was projected into a 2-dimensional vector space constructed by the Principal Component Analysis (PCA) for each speaker [19]. The results clearly showed that the distribution of mean MFCC vectors of dialogue speech was closer to the center of the distribution of all the phonemes than the distribution of read speech. In other words, the size of spectral space for the phonemes in spontaneous speech was smaller compared to that of read speech.

4.3. Reduction ratio of the distribution of phonemes

In order to quantitatively analyze the reduction of the distribution of phonemes, Euclidean norms/distances between the mean vector of each phoneme and the center of the distribution of all phonemes (the vector averaged over all the phonemes) were calculated, and the ratio of the distance for spontaneous speech (presentations and dialogues) to that of read speech was calculated for each phoneme as follows:

$$red_p(X) = \frac{\|\mu_p(X) - \mathrm{Av}[\mu_p(X)]\|}{\|\mu_p(R) - \mathrm{Av}[\mu_p(R)]\|} \quad (1)$$

Here $\mu_p(X)$ is the mean vector of a phoneme p uttered with a speaking style X, $\mu_p(R)$ is the mean vector of a phoneme p of read speech, and Av indicates the averaged value.

Results using the mean MFCC vector of the second state of the HMM with a single Gaussian mixture as the mean vector for each phoneme are shown in Figure 4.

The figure shows the reduction ratios $red_p(X)$ averaged over all the speakers, separately for AP, EP, and dialogues. /N:/ and /ch/, which rarely occurred in the utterances were not used in this analysis. The condition of $red_p(X) = 1$ is indicated by a thick line. The dialogues include interviews concerning AP and EP, task dialogues, and free dialogues. Results in the figure show the reduction of the MFCC space for almost all the phonemes in the three speaking styles, and this is most significant for dialogue utterances.

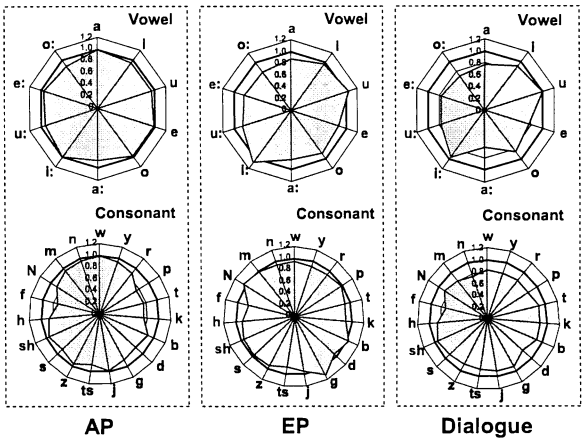

Figure 4: The reduction ratio of the vector norm between each phoneme and the phoneme center in spontaneous speech to that in read speech.

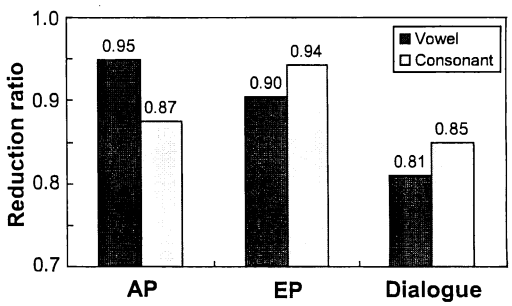

Figure 5: Mean reduction ratios of vowels and consonants for each speaking style.

Figure 5 shows mean reduction ratios for vowels and consonants, respectively, for each speaking style. These results show that the reduction of the distribution of phonemes in the MFCC domain in comparison with that of read speech is observed for all the speaking styles, and most significantly for dialogue speech.

4.4. Reduction of distances between phonemes

In the previous section, the reduction of the MFCC space was measured by the ratio of the distance between each phoneme and the phoneme center in spontaneous speech to that in read speech. In this section, the reduction of the cepstral distance between each phoneme pair is measured. The Euclidean distance using the mean MFCC vector of each phoneme and the Mahalanobis distance, which takes into account the variances, were measured. The definition of the Mahalanobis distance $D_{ij}(X)$ between phoneme i and j spoken with speaking style X can be

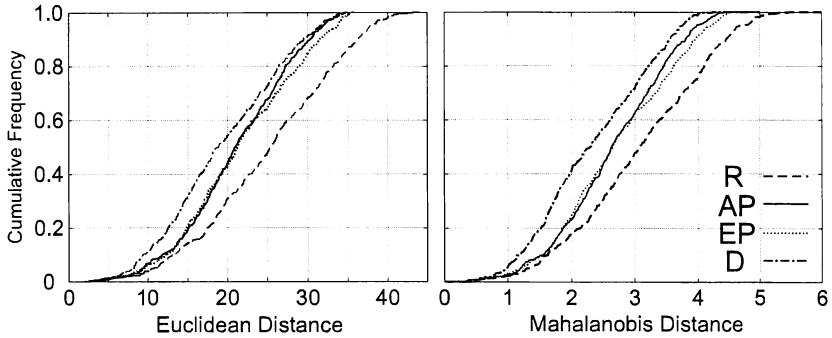

Figure 6: Distribution of distances between phonemes.

written as follows.

$$D_{ij}(X) = \sqrt{\frac{K\sum_{k=1}^{K}(\mu_{ik}(X) - \mu_{jk}(X))^2}{\sum_{k=1}^{K}\sigma_{ik}^2(X) + \sum_{k=1}^{K}\sigma_{jk}^2(X)}} \quad (2)$$

Where, K is the dimension of an MFCC vector ($K = 12$). $\mu_{ik}(X)$ and $\sigma_{ik}^2(X)$ are the kth dimensional elements of the mean and the variance vector of MFCC for phoneme i uttered with a speaking style X. In the case of the Euclidean distance between phonemes i and j, the denominator in the above formula (2) is set to a constant value.

Five males and five females were randomly selected from the CSJ for this experiment. The total number of phoneme samples for each speaking style was 45,242 (read speech), 80,095 (AP), 55,102 (EP), or 56,583 (dialogues). The read speech set in the CSJ includes various kinds of "reading transcriptions" and "reading novels including dialogues". The dialogue set includes variations of "interview" and "free dialogue". Therefore, speech materials of read speech and dialogues for this experiment were selected so as to represent as many variations of speaking styles as possible. Speech materials of AP and EP were randomly selected from the test-set data of CSJ designed for speech recognition experiments.

Figure 6 shows the cumulative frequency of distances between phonemes for each speaking style. The left-hand side of the figure shows the case using the Euclidean distance, whereas the right-hand side shows the case using the Mahalanobis distance. The horizontal axis indicates the Euclidean or the Mahalanobis distance, and the vertical axis indicates the cumulative frequency. These results clearly show that the distance between phonemes decreases as the spontaneity of the utterances increases (D $\gg$ EP > AP $\gg$ R). The Wilcoxon's rank order test with a significance level of p-value ≤ 0.01 shows that the distributions of each speaking style are statistically different from each other, except between AP and EP.

4.5. Relationship between phoneme distances and phoneme recognition performance

Differences in the size of the distribution of between-phoneme distances are expected to be related to phoneme recognition performance for various speaking styles. This section investigates the relationship between between-phoneme distances and phoneme recognition accuracy using utterances by many speakers. Mono-phone HMMs with a single Gaussian mixture for phoneme recognition were trained for each speaking style, using utterances by 100 males and 100 females for AP and 150 males and 150 females for EP. These speakers were randomly selected from the CSJ, and the total number of phoneme samples were approximately two million for AP and EP, respectively. A 38-dimensional feature vector was used as the acoustic feature. The same data as used in Section 4.4 were used for the evaluation experiment. A phoneme network with di-phone probabilities was used as a language model for recognition. The insertion penalty was optimized for each speaking style.

Figure 7 shows the relationship between the mean phoneme distance and the phoneme recognition accuracy. The left-hand side of the figure shows the case using Euclidean distance and the right-hand side shows the case using Mahalanobis distance as the distance between phonemes for each speaking style. Correlation coefficients between the mean phoneme distance and the phoneme recognition accuracy are 0.93 in the case using Euclidean distance and 0.97 in the case using Mahalanobis distance. The lines in Figure 7 indicate the regression over the four points. These results clearly show a strong correlation between mean phoneme distance and phoneme accuracy. This means that the phoneme recognition accuracy can be estimated by the mean phoneme distance. That is, the reduction of the distances between phonemes is a major factor contributing to the degradation of spontaneous speech recognition accuracy. It can also be concluded that the relationship between the phoneme distance and the phoneme recognition accuracy becomes slightly more significant if the variances of phoneme spectra are also taken into account.

5. Comparison between Language Models of Written Text and Spontaneous Speech

5.1. Experimental Conditions

Language models built by written text and spontaneous speech were compared, using Mainichi newspaper (NP) and news commentary (NC) corpora as written text corpora, and AP, EP and dialogues (D) in the CSJ as spontaneous speech corpora [21]. The NC has features between written text and spontaneous speech, since it consists of transcription of utterances spoken based on prepared text. The total number of words and the vocabulary size for each corpus are shown in Table 3. All the texts were segmented into morphemes by the morphological analy-

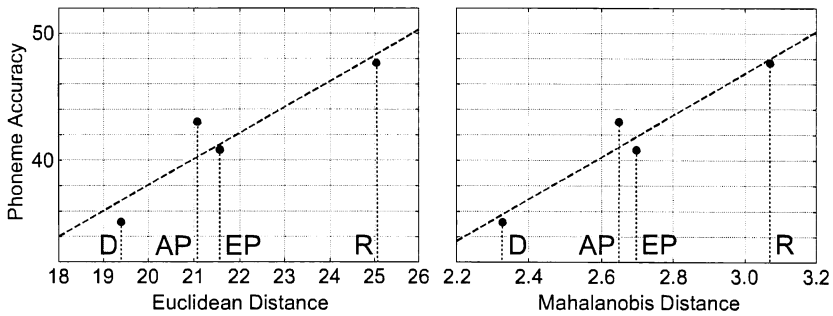

Figure 7: Relationship between phoneme distances and phoneme recognition accuracy.

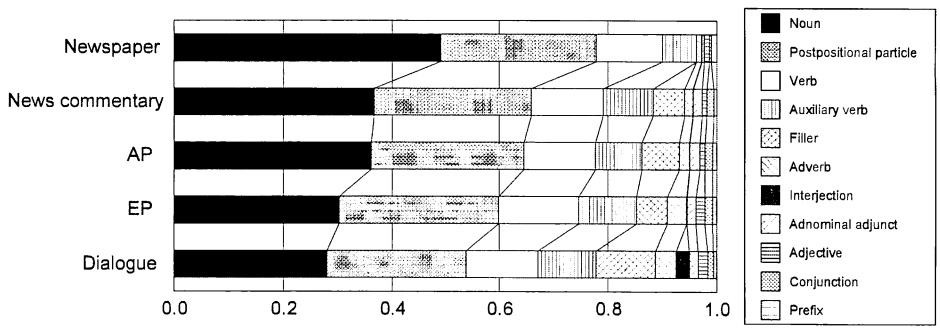

Figure 8: Observation frequency of each part-of-speech in each corpus.

Table 3: Total number of words used for training and vocabulary size for each language model (NP: newspaper, NC: news commentary, AP: academic presentation, EP: extemporaneous presentation, D: dialogue)

	Number of words	Vocabulary size
NP	5,157,852	30,000
NC	927,207	18,044
AP	3,294,234	29,674
EP	3,671,259	30,000
D	121,541	6,266

sis program, Chasen (Ver 2.3.3). Orthographical errors in the analysis results were automatically corrected by Chawan (Ver 2.06). Words having not only different characters but also different pronunciations or part-of-speeches were stored as different entries in the dictionary for language modeling. The symbol ⟨sp⟩ was given to each silent period with a duration of 200-500 ms and a comma in the text, and the symbol ⟨sil⟩ was given to each silent period with a duration longer than 500 ms and a period in the text in the language modeling process. The number of part-of-speeches used in this experiment was 15. The toolkit Palmkit (Ver 1.0.30) was used for language model training and evaluation.

5.2. Part-of-Speech Observation Frequency

Figure 8 indicates variation of the observation frequency of part-of-speech in each corpus. There is a significant difference of the part-of-speech frequency between written text and spontaneous speech. The frequency of nouns is much higher in the newspaper corpus than in the spontaneous speech, and the frequency of fillers is much higher in the dialogue than in the news commentary and presentation corpus. A supplementary analysis of word bigrams show that the frequency of noun-noun concatenation is significantly higher in the newspaper corpus. This means that one of the reasons for the high frequency of nouns in the newspaper corpus is the fact that frequently occurring compound nouns are split into constituent nouns.

5.3. Perplexity

Trigrams were built as statistical language models for each corpus, and test-set perplexity and out-of-vocabulary (OOV) rate were measured for every combination of the corpora. The perplexities and OOV rates are shown in Tables 4 and 5, respectively. Since vocabulary sizes are different for each corpus and the dialogue corpus has not only a significantly smaller vocabulary size than others but also a large OOV rate, strict comparison cannot be made between different corpora. Nevertheless, it can be clearly observed that, even if each language model is built using the same corpus as the test set (diagonal elements in Table 4), perplexity for spontaneous speech, such as presentations and dialogues, is roughly five times larger than that for written text, such as newspapers.

The perplexity matrix in Table 4 can be considered approximately representing the distances between the language models for different corpora. In order to visualize the distances between the language models, the perplexity matrix is symmetrized by using equation (3) and then used for Multidimensional Scaling.

Table 4: Test-set perplexities for every combination of the corpora

Test set	Language model				
	NP	NC	AP	EP	D
NP	**21.06**	293.27	571.08	440.92	487.92
NC	148.70	**69.74**	210.38	187.02	343.49
AP	406.29	204.02	**100.88**	192.37	382.04
EP	321.19	258.16	171.35	**93.03**	190.03
D	695.08	644.16	328.74	169.15	**111.04**

Table 5: OOV rates for every combination of the corpora

Test set	Language model				
	NP	NC	AP	EP	D
NP	**1.32**	6.11	6.12	4.70	25.83
NC	2.23	**0.90**	2.84	1.68	22.45
AP	9.21	6.46	**1.88**	3.41	19.16
EP	7.50	5.63	3.28	**1.08**	16.05
D	6.70	11.49	3.98	4.71	**5.06**

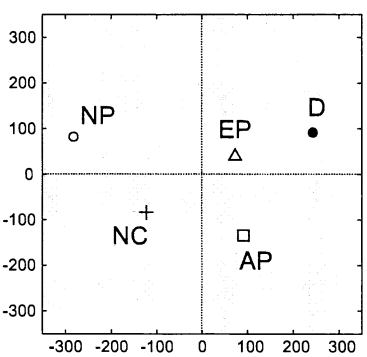

Figure 9: Relationships between the language models derived from the perplexity matrix.

In equation (3), a_{ij} is the i-th row, j-th column component of the perplexity matrix, d_{ij} is the i-th row, j-th column component of the (symmetrized) distance matrix, and T is the number or corpora (T=5).

$$d_{ij} = \sum_{i=1}^{T} \sum_{j=1}^{T} \frac{(a_{ij} + a_{ji}) - (a_{ii} + a_{jj})}{2} \qquad (3)$$

Figure 9 shows the relationships between the language models. Newspaper text and dialogue are situated at two extreme positions, and presentations and news commentary are situated in between.

Recognition experiments using a common acoustic model and different language models trained for each corpus were conducted to investigate the relationship between perplexity (diagonal elements in Table 4) and word accuracy. Experimental results show that they have a high correlation of -0.98.

6. Conclusion

In order to increase recognition accuracy for spontaneous speech, it is necessary to build acoustic and language models using spontaneous speech corpora. It was found through our recognition experiments for spontaneous presentations in the Corpus of Spontaneous Japanese (CSJ), that recognition accuracy increases as training data size increases, even up to 510 hours or 6.84M words for both acoustic and language model training. This indicates that spontaneous speech is so variable that it needs a huge corpus to encompass the variations. However, it is impossible to collect a huge corpus for every new application. Therefore, it is important to clarify general features of spontaneous speech and establish a mechanism for adapting a task-independent model to a specific task using task-specific features [3, 22, 23, 24].

By comparing spontaneous speech and speech reading the transcription of that spontaneous speech, it was clarified that the spectral distribution of spontaneous speech is significantly reduced compared to that of read speech. Although this was true for all the spontaneous speech analyzed in this paper, that is, academic presentations (AP), extemporaneous presentations (EP), and dialogues, the reduction was most significant for dialogues, which are obviously more spontaneous than the other styles. It has also been found that the more spontaneous the speech, the smaller the distances between phonemes become; and, that there is a high correlation between the mean phoneme distance and the phoneme recognition accuracy. In summary, spontaneous speech can be characterized by the reduction of spectral space in comparison with that of read speech, and this is one of the major factors contributing to the decrease in recognition accuracy.

As for linguistic characteristics of spontaneous speech, trigram-based language models were built for various corpora, including newspaper text as a written text corpus, and test-set perplexities were compared. The results show that the perplexity of spontaneous speech, including academic as well as extemporaneous presentations and dialogues, is significantly higher than that of written text, and the perplexity of news commentary, which is reading written text, is located between written text and spontaneous speech. It was also found that the frequency of fillers is significantly higher for dialogues compared with news commentaries or presentations.

Our future research includes analysis over wider ranges of spontaneous speech using utterances other than those included in the CSJ. Broadening speech recognition applications depends crucially on raising the recognition performance of spontaneous speech. Although we have clarified spectral reduction and its effects on spontaneous speech recognition, it is not yet clear how we can use these results for improving recognition performances. Creating methods for adapting acoustic models to spontaneous speech based on the results obtained in this research is one of our future targets.

The perplexity and OOV for language models of spontaneous speech is significantly higher than that for written text. This is due to the fact that spontaneous speech frequently includes ungrammatical phenomena and linguistic variations, including repetitions and repairs. In order to cover these variations, it is crucial to collect a huge spontaneous speech corpus to train the language models. However, to build a spontaneous corpus, we need to record actual speech, manually transcribe and segment it into sentences and words, and annotate various information. Since this is a much more labor-intensive and costly process than that needed for building a read speech corpus, existing spontaneous speech corpora are much smaller than those of written text or read speech. How to efficiently collect and utilize spontaneous speech corpora for training language models is one of the most important issues for improving spon-

taneous speech recognition performance. How to incorporate filled pauses, repairs, hesitations, repetitions, partial words, and disfluencies still poses a big challenge.

The large-scale spontaneous speech corpus, CSJ, used in the experiments reported in this paper, is stored with XML format in a large-scale database system developed by the 21st Century COE (Center of Excellence) program "Framework for Systematization and Application of Large-scale Knowledge Resources" at Tokyo Institute of Technology so that the general population can easily access and use it for research purposes [25]. We hope international collaboration in building large-scale spontaneous speech corpora as well as analysis and modeling of spontaneous speech based on the corpora will advance the progress of speech recognition technology.

7. References

[1] Furui, S.: Recent advances in spontaneous speech recognition and understanding. Proc. IEEE Workshop on SSPR, Tokyo (2003) 1–6.

[2] Furui, S.: Toward spontaneous speech recognition and understanding. In: Pattern Recognition in Speech and Language Processing, Chou, W. and Juang, B.-H., Eds., CRC Press, New York (2003) 191–227.

[3] Shinozaki, T., Hori, C. and Furui, S.: Towards automatic transcription of spontaneous presentations. Proc. Eurospeech, Aalborg, Denmark (2001) 491–494.

[4] Sankar, A., Gadde, V. R. R., Stolcke, A. and Weng, F.: Improved modeling and efficiency for automatic transcription of broadcast news. Speech Communication, vol.37 (2002) 133–158.

[5] Gauvain, J.-L. and Lamel, L.: Large vocabulary speech recognition based on statistical methods. In: Pattern Recognition in Speech and Language Processing, Chou, W. and Juang, B.-H., Eds., CRC Press, New York (2003) 149–189.

[6] Evermann, G. et al.: Development of the 2003 CU-HTK conversational telephone speech transcription system. Proc. IEEE ICASSP, Montreal (2004) I-249–252.

[7] Schwartz, R. et al.: Speech recognition in multiple languages and domains: the 2003 BBN/LIMSI EARS system. Proc. IEEE ICASSP, Montreal (2004) III-753–756.

[8] van Son, R.J.J.H. and Pols, L.C.W.: An acoustic description of consonant reduction. Speech Communication, vol.28, no.2 (1999) 125–140.

[9] Duez, D.: On spontaneous French speech: aspects of the reduction and contextual assimilation of voiced stops. J. Phonetics, vol.23 (1995) 407–427.

[10] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation. Proc. IEEE Workshop on SSPR, Tokyo (2003) 7–12.

[11] Maekawa, K., Kikuchi, H. and Tsukahara, W.: Corpus of Spontaneous Japanese: design, annotation and XML representation. Proc. International Symposium on Large-scale Knowledge Resources, Tokyo (2004) 19–24.

[12] Uchimoto, K., Nobata, C., Yamada, A., Sekine, S. and Isahara, H.: Morphological analysis of the Corpus of Spontaneous Japanese. Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo (2003) 159–162.

[13] Venditti, J.: Japanese ToBI labeling guidelines. OSU Working Papers in Linguistics, vol.50 (1997) 127–162.

[14] Maekawa, K., Kikuchi, H., Igarashi, Y. and Venditti, J.: X-JToBI: an extended J-ToBI for spontaneous speech. Proc. ICSLP, Denver, CO (2002) 1545–1548.

[15] Kawahara, T., Nanjo, H., Shinozaki, T. and Furui, S.: Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo (2003) 135–138.

[16] Shinozaki, T. and Furui, S.: Analysis on individual differences in automatic transcription of spontaneous presentations. Proc. IEEE ICASSP, Orlando (2002), I-729–732.

[17] Ichiba, T., Iwano, K. and Furui, S.: Relationships between training data size and recognition accuracy in spontaneous speech recognition. Proc. Acoustical Society of Japan Fall Meeting (2004) 2-1-9. (in Japanese)

[18] Ueberla, J.: Analysing a simple language model – some general conclusion for language models for speech recognition. Computer Speech & Language, vol.8, no.2 (1994) 153–176.

[19] Nakamura, M., Iwano, K. and Furui, S.: Analysis of spectral space reduction in spontaneous speech and its effects on speech recognition performances. Proc. INTERSPEECH 2005 (2005), 3381–3384.

[20] Furui, S., Nakamura, M., Ichiba, T. and Iwano, K.: Analysis and recognition of spontaneous speech using *Corpus of Spontaneous Japanese*. Speech Communication, vol.47, no.1–2 (2005) 208–219.

[21] Nakamura, M., Iwano, K. and Furui, S.: Analysis of linguistic characteristics of spontaneous speech in comparison with read speech. Proc. Acoustical Society of Japan Fall Meeting (2005), 3-6-10. (in Japanese)

[22] Nanjo, H. and Kawahara, T.: Unsupervised language model adaptation for lecture speech recognition. Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo (2003) 75–78.

[23] Lussier, L., Whittaker, E. W. D. and Furui, S.: Combinations of language model adaptation methods applied to spontaneous speech. Proc. Third Spontaneous Speech Science & Technology Workshop, Tokyo (2004), 73–78.

[24] Shinozaki, T. and Furui, S.: Spontaneous speech recognition using a massively parallel decoder. Proc. Interspeech-ICSLP, Jeju, Korea, vol.3 (2004) 1705–1708.

[25] Furui, S.: Overview of the 21st century COE program "Framework for Systematization and Application of Large-scale Knowledge Resources". Proc. International Symposium on Large-scale Knowledge Resources, Tokyo (2004) 1–8.