

Comparison of Effects of Acoustic and Language Knowledge on Spontaneous Speech Perception/Recognition between Human and Automatic Speech Recognizer

Norihide Kitaoka, Masahisa Shingu, Seiichi Nakagawa

Department of Information and Computer Sciences
Toyohashi University of Technology, Japan
{kitaoka, shingu, nakagawa}@slp.ics.tut.ac.jp

Abstract

An automatic speech recognizer uses acoustic knowledge and linguistic knowledge. In large vocabulary speech recognition, acoustic knowledge is modeled by hidden Markov models (HMM), linguistic knowledge is modeled by N-gram (typically bi-gram or trigram), and these models are stochastically integrated. It is thought that humans also integrate acoustic and linguistic knowledge of speech when perceiving continuous speech. Automatic speech recognition with HMM and N-gram is thought to roughly model the process of human perception.

Although these models have drastically improved the performance of automatic speech recognition of well-formed read speech so far, they cannot deliver sufficient performance on spontaneous speech recognition tasks because of various particular phenomena of spontaneous speech.

In this paper, we conducted simulation experiments of N-gram language models by combining human acoustic knowledge and instruction of local context and assured that using two words neighboring the target word was enough to improve the performance of recognition when we could use only local information as linguistic knowledge. We also assured that coarticulation affected the perception of short words.

We then compared some language models on speech recognizer. We calculated acoustic scores with HMM and then linguistic scores calculated from a language model were added. We obtained 37.5% recognition rate only with acoustic model, whereas we obtained 51.0% with both acoustic and language models, thus the relative performance improvement was 36%. On the other hand, we obtained a 16.5% recognition rate only with the language model, so the acoustic model improved the performance relatively 209%. The performance of the language model on spontaneous speech is almost equal to that on read speech and thus, the improvements of the acoustic models is more effective than that of the language model.

1. Introduction

The performance of large vocabulary continuous speech recognition (LVCSR) is sufficient for read speech, and the systems are practically used in dictation systems, broadcasting systems [1, 2], etc. On the contrary, recognition of dialog or lecture speech is also demanded, but the performance is still poor for these tasks. There are many reasons for this degradation: much hesitation, repetition, fast utterance speed, large coarticulations, etc.

In large vocabulary speech recognition, hidden Markov models (HMM) and N-gram (typically with small N such as bi-gram (2) or trigram (3)) models are used as acoustic models and language models, respectively. An HMM statistically

models local acoustic features of a sub-word such as a phone or a syllable, usually including short-term coarticulation with just preceding or succeeding sub-words, and N-gram model statistically models local linguistic properties such as a conditional probability when given preceding $N - 1$ words. Some studies have been performed to reveal the reason for the performance degradation. Nanjo et al. [3, 4] developed a method to diagnose recognition errors, that is, which model was the reason for misrecognition. Shinozaki et al.[5] analyzed the reason for individual differences in speech recognition performance.

Humans must use more global information. Some efforts have been made to model such information, but general method to use such global information has not been developed so far. In this paper, we show simulation experiments of N-gram language models by combining human acoustic knowledge and instruction of local context and discuss the performance of such local linguistic models.

2. Perception experiments by human

N-gram can be regarded as a model which to model the statistics of language by the following approximation:

$$P(w_i|W_1^I) = P(w_i|w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_I) \quad (1)$$

$$\approx P(w_i|w_{i-N+1}, \dots, w_{i-1}). \quad (2)$$

Thus, a long context including preceding and succeeding words is approximated by short (typically 2 word length) one including only preceding words. Although this approximation may be too radical, this modeling achieved good performance on read speech recognition in combination with the HMM acoustic model. We evaluated some variation of context approximation by simulation using human acoustic perceptive ability.

2.1. Setup

We evaluated word perception performance of humans when some word context was given.

We used two sets of speech as test sets. Speech from 4 lectures (A01M0074, A01M0035, A01M0007, A05M0031) included in the Corpus of Spontaneous Japanese (CSJ) were used as spontaneous speech. These were oral presentations at conference meetings related to speech in Japan. We randomly chose 50 words as targets from each lecture regardless of the part of speech except fillers and unknown words which were not included in the vocabulary used in Section 3. This set contains 200 words. We also used read speech from JNAS Japanese corpus[6]. We selected 100 words from the corpus as the same

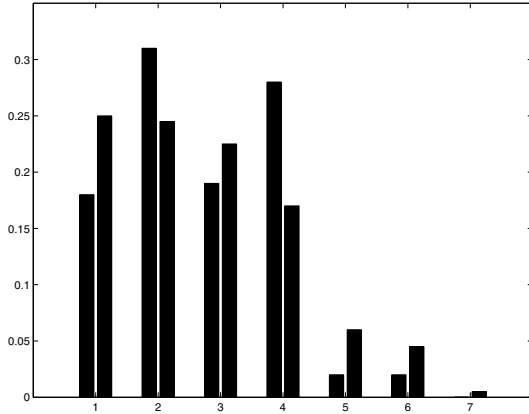


Figure 1: Histogram of length of target words in syllables. (Left: read speech, right: spontaneous speech)

way in the case of spontaneous speech. The histogram of the word length is shown in Figure 1.

Each target word was manually segmented with the word(s) in the context below:

No context: Target word only;

2 preceding words: Target word along with two preceding ones;

1 word each: Target word along with one preceding and one succeeding words;

2 words each: Target word along with two preceding and two succeeding words.

Testees listened to each segment under the condition that they knew the word sequence context (no *a priori* knowledge was taught in the case of ‘No context’, of course), and they wrote what the target word was heard as.

Target words in each corpus were arranged at random to avoid the effects of information from other targets and their local contexts. Testees could listen to each word repeatedly, but were prohibited from going back words they had listened to before.

Testees were all students of a master or doctor course in a field related to speech processing because they should have some knowledge of the field of the lectures, such as technical terms, to simulate the language models adapted to a lecture speech recognition system.

Since same target words were used under all conditions, this might have impacted the results. To avoid this, we were careful about the order and intervals between tests.

2.2. Results

We performed the test introduced in Section 2.1. We made some corrections. Some homonym pairs were both considered as the same words because of their similar meanings. We also considered some variation of pronunciations. For example, a syllable /zu/ appearing in some particular words is sometimes pronounced as /tsu/ and this variation does not affect understanding of the words, so the description of such words with /tsu/ is judged as correct.

Table 1 shows the results. We can see the effect of word length on perception performance. In both cases of read speech and spontaneous speech, the performance under 2 preceding

Table 1: Human word perception rate with various word contexts[%]

(a) Read speech						
word context		word length (no. syllables)				
preceding	succeeding	1	2	3	≥ 4	ave.
0	0	57.4	72.0	92.1	93.8	80.2
2	0	96.3	98.9	97.4	97.9	97.8
1	1	97.2	98.9	100.0	99.5	99.0
2	2	98.1	100.0	99.1	99.5	99.3

(b) Spontaneous speech						
word context		word length (no. syllables)				
preceding	succeeding	1	2	3	≥ 4	ave.
0	0	52.7	55.8	75.6	87.8	68.4
2	0	95.3	87.4	87.4	91.1	90.4
1	1	94.7	93.5	94.8	96.7	95.0
2	2	96.3	95.2	97.0	95.8	96.3

Table 2: Perplexity for each word length by unigram (corresponding to no context) and trigram language model (corresponding to 2 preceding word context).

(a) Read speech					
LM	word length (no. syllables)				
	1	2	3	≥ 4	ave.
unigram	506	4703	16116	10426	5133
trigram	41	163	211	1950	296

(b) Spontaneous speech					
LM	word length (no. syllables)				
	1	2	3	≥ 4	ave.
unigram	704	1488	8168	7436	2841
trigram	171	370	2101	1763	698

words condition was much better than in no context condition. The smaller the number of syllables in a word was, the greater the performance improvement was. Short words were difficult to perceive with their small acoustic information, but the constraint of linguistic information (in this case, a trigram) was used effectively.

Table 2 (b) shows the perplexity for each word length with language model for spontaneous speech described in Section 3.1. Most of particles in Japanese consist of one syllable and the perplexities of particles tend to be small when using a bigram or trigram. Prediction performance of target words only with word context information (that is, without acoustic information) is shown in Table 3. Trigram model almost corresponds to 1-word each context condition and we can easily find a correlation between perplexities for trigram in Table 2 and perception performance of 1-word each context condition in Table 3.

The performances under 1 word each condition and 2 preceding words condition were almost the same for read speech. On the contrary, the performance under 1 word each condition was significantly superior to that under 2 preceding words condition even though these two contexts had almost the same linguistic information quantities. We assumed that the difference between read speech and spontaneous speech was caused by coarticulation and we removed the effect of linguistic con-

Table 3: Human word prediction rates only with word contexts (spontaneous speech)[%]

word context		word length (no. syllables)				
preceding	succeeding	1	2	3	≥ 4	ave.
1	1	24.9	11.1	2.5	2.0	10.1
2	2	41.7	25.7	6.7	9.9	21.0

Table 4: Human word perception rates with/without syllable/word contexts [%]

(a) Read speech			
context		word length (no. syllables)	
preceding	succeeding	1	2
0	0	57.4	72.0
1-syl.	1-syl.	75.9	87.6
1-word	1-word	97.2	98.9

(b) Spontaneous speech			
context		word length (no. syllables)	
preceding	succeeding	1	2
0	0	52.7	55.8
1-syl.	1-syl.	71.3	89.1
1-word	1-word	94.7	93.5

text. Table 4 shows the comparison of perception performances of one-syllable and two-syllable words with/without preceding and succeeding syllables. The perception performance of words in spontaneous speech got almost the same as that in read speech with one-syllable each context especially for two-syllable words. The context scarcely had linguistic information, so this improvement was almost only the result of acoustic information related to coarticulation. This result suggests that acoustic modeling robust against the large coarticulation is important for spontaneous speech.

When assuming that the coarticulation of acoustics affects only the acoustics of the neighboring syllable, the difference between 1-syllable each condition and 1-word each condition is only the quantity of linguistic information. Performance differences between these two conditions are almost same for read speech and spontaneous speech. This large difference was brought by linguistic information. In Table 1, we can also find that the improvement of linguistic information doesn't improve the perception performance when comparing 1 word each condition and 2 words each condition. Only with linguistic information, we could find drastic improvement of prediction performance when context got longer in Table 3, but 1-word each linguistic information was sufficient for short words and acoustic information was relatively larger than linguistic information for long words. So the improvement of linguistic information does not always improve the perception performance.

3. Recognition experiments by automatic speech recognizer

We tested an automatic speech recognizer under almost the same conditions under which perception experiments described in Section 2 was performed.

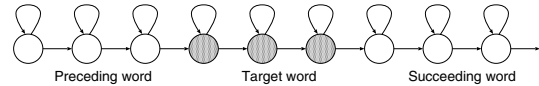


Figure 2: HMM to calculate acoustic score for target word.

3.1. Experimental conditions

We chose the same target words used in Section 2 and manually segmented each word with a preceding word and a succeeding word. We calculated the acoustic likelihood for each word in vocabulary with the concatenated word HMM as shown in Figure 2. Then we added the linguistic score with appropriate weight to the acoustic score to obtain the total score.

Speech data were sampled with a sampling frequency of 12 kHz, and the signal was pre-emphasized by a factor of 0.97. A Hamming window length of 25 ms was applied and shifted with the step of 10 ms. 38 dimensional feature vectors were used including 12 dimensional MFCCs, their first and second deviation coefficients and the first and second deviations of log power were used. MFCCs were derived through 22 dimensional Mel filter banks from speech by a tool of HTK ver.3.0 [7].

For read speech, 114 Japanese context-independent syllable HMMs were trained using 27992 utterances read by 175 male speakers (JNAS corpus). For spontaneous speech, we adapted the HMMs to lecture speech (a part of CSJ corpus consisting of 204 lectures presented by male speakers). Each continuous HMM had 5 states, and 4 of them had pdfs of output probability. Each pdf consisted of four Gaussians with a full-covariance matrix.

We used two trigram language models. The models for read speech was trained from the text of 45 months Mainichi Japanese newspaper with 90 million words. For spontaneous speech, we used the model trained at Kyoto University using lecture speech transcriptions provided as trial version with CSJ corpus.

3.2. Results

Recognition results by speech recognizer with/without language models are shown in Table 5. When using only acoustic models, the recognition rates were 60.0% for read speech and 37.5% for spontaneous speech. No improvements were obtained when using the unigram language models. The trigram language models improved the recognition rates to 64.0% and 51.0%, respectively.

Table 6 shows recognition results with/without acoustic models. This can show the prediction ability of the language models. 20.0% and 16.5% of all the target words for read speech and for spontaneous speech, respectively, were correctly predicted as the first candidate and was improved to 64.0% and 51.0%, respectively, by integrating with acoustic model.

This prediction performance was superior to that of human described as that on 1 word each condition in Table 3. In general, human can predict the words more accurately if he/she is given longer contexts, but the performance of N -gram models is no longer improved with larger N [8]. That is, even though the prediction performance of 2 words each condition was much better than that of 1 word each condition, we cannot expect the improvement of the performance of N -gram model with larger N . Even with the long context, improvement of human perception performance was not so large (as described in Table 1). This suggests that $N = 3$ for a N -gram model is sufficient.

Table 5: Comparison of recognition results with acoustic score (AS) only and with acoustic score + language score.
Read speech

		AS only		
		correct	incorrect	
AS+LS	correct	51.0	1.0	52.0
	incorrect	9.0	39.0	48.0
		60.0	40.0	

(a) unigram

		AS only		
		correct	incorrect	
AS+LS	correct	57.0	7.0	64.0
	incorrect	3.0	33.0	36.0
		60.0	40.0	

(b) trigram

Spontaneous speech

		AS only		
		correct	incorrect	
AS+LS	correct	28.0	8.5	36.5
	incorrect	9.5	54.0	63.5
		37.5	62.5	

(a) unigram

		AS only		
		correct	incorrect	
AS+LS	correct	30.0	21.0	51.0
	incorrect	7.5	41.5	49.0
		37.5	62.5	

(b) trigram

Table 6: Comparison of recognition results with language score (LS) only and with acoustic score + language score.
Read speech

		LS only		
		correct	incorrect	
AS+LS	correct	1.0	51.0	52.0
	incorrect	0.0	48.0	48.0
		1.0	99.0	

(a) unigram

		LS only		
		correct	incorrect	
AS+LS	correct	18.0	46.0	64.0
	incorrect	2.0	34.0	36.0
		20.0	80.0	

(b) trigram

Spontaneous speech

		LS only		
		correct	incorrect	
AS+LS	correct	1.0	35.5	36.5
	incorrect	1.0	62.5	63.5
		2.0	98.0	

(a) unigram

		LS only		
		correct	incorrect	
AS+LS	correct	13.5	37.5	51.0
	incorrect	3.0	46.0	49.0
		16.5	83.5	

(b) trigram

4. Conclusion

We simulated N-gram language models using human acoustic knowledge and instruction of local context. We assured that using two words immediately before the target word as context sufficed to obtain good recognition performance when using local information as linguistic knowledge. We also assured that the coarticulation affected human perception of short words.

We compared some language models on the speech recognizer. We calculated the acoustic score of a word with HMM and added linguistic scores calculated from language models to the acoustic scores. We obtained 60.0% and 37.5% recognition rate only with acoustic models for read speech and spontaneous speech, respectively, whereas we obtained 64.0% and 51.0% with both acoustic and language models, respectively, thus the relative performance improvement was 6.7% and 36%. We obtained 20.0% and 16.5% recognition rate only with the language model, so the performance of language models for spontaneous speech was little inferior to that for read speech. The prediction ability of the language model was not inferior to that of human, but the recognition performance of the recognizer was much worse than the human perception performance. Thus, we can expect that the improvements of the acoustic models is more effective than that of the language model.

5. References

[1] Imai, T., Kobayashi, A., Sato, S., Tanaka, H., Ando, A., "Progressive 2-Pass decoder for real-time broadcast news

captioning", in proceedings of ICASSP-2000, vol. 3, pp. 726–729, 2000.

[2] Nguyen, L., Guo, X., Schwartz, R., Makhoul, J., "Japanese broadcast news transcription", in proceedings of ICSLP-2002, pp. 1749–1752, 2002.

[3] Nanjo, H., Lee, A., Kawahara, T., "Automatic diagnosis of recognition errors in large vocabulary continuous speech recognition systems" in proceedings of ICSLP-2000, vol. 2, pp. 1027–1030, 2000.

[4] Nanjo, H., Kato, K., Mimura, M., Lee, A., Kawahara, T., "Diagnosis and evaluation of various LVCSR systems", IPSJ SIG notes, SLP-2000–31–11, pp. 73–80, 2000.

[5] Shinozaki, T., Furui, S., "Assessment of the performance of automatic spontaneous speech recognition system using human recognition performance", The 2002 Autumn meeting of the Acoustical Society of Japan, vol. 2–8–13, pp. 87–88, 2002.

[6] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuo, T., Kobayashi, T., Shikano, K., Itahashi, S., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research", The Journal of the Acoustical Society of Japan(E), Vol.20, pp.199-206, 1999.

[7] Young, S., Kershov, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., "The HTK Book", 2000.

[8] Owens, M., Kruger, A., Donnelly, P., Smith, P. D., Ming, J., "A missing-word test comparison of human and statistical language model performance", in proceedings of Eurospeech-99, vol. I, pp. 145–148, 1999.