



# Improving the Readability of Class Lecture ASR Results using a Confusion Network

Yasuhisa Fujii, Kazumasa Yamamoto, Seiichi Nakagawa

Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

{fujii, kyama, nakagawa}@slp.cs.tut.ac.jp

## Abstract

This paper presents a method for improving the readability of Automatic Speech Recognition (ASR) results for classroom lectures. Most of the previous research on improving the readability of recognition results focused mainly on manually transcribed texts, and not ASR results. Due to the presence of a large number of domain-dependent words and the casual presentation style, even state-of-the-art recognizers yield a 30-50% word error rate for speech in classroom lectures. Thus, a method for improving the readability of ASR results needs to be robust against recognition errors. In this paper, we propose a novel method for improving the readability based on a machine translation model that uses a confusion network representing multiple hypotheses of the ASR results to achieve robustness against recognition errors. Experimental results show that the proposed method outperforms the baselines in both automatic and manual evaluations.

**Index Terms:** improving readability, confusion network, automatic speech recognition, classroom lecture speech

## 1. Introduction

The availability of an audio transcription of a speech allows the content of the speech to be more easily understood. In particular, classroom lectures, which are the focus of this paper, benefit from transcriptions because these transcriptions can assist the hearing impaired and can also be used in downstream processing, such as summarization [1], indexing [2], browsing systems [3], and so on. As a result, much research is currently underway on transcribing these lectures [4, 5, 6].

However, automatic speech recognition of classroom lectures is quite difficult due to the presence of a large number of domain-dependent words and the spontaneity of the speaker. Thus, state-of-the-art recognizers typically achieve a Word Error Rate (WER) of around 30-50% [4, 5, 6].

Furthermore, the recognition results from existing Automatic Speech Recognition (ASR) systems are not easily understood by humans even in the case of perfect speech recognition results, because the speech used in classroom lectures contains many ill-formed utterances with filled pauses, restarts, repetitions, deletion of prepositions, and so on. Thus, before making transcripts available to users, the readability thereof needs to be improved to assist the readers in understanding the contents of the lecture material. In this context, extensive research is currently underway on paraphrasing and correcting recognition results [7, 8, 9]. Shitaoka et al. formulated the problem as a framework of machine translation and applied a statistical method to transform spoken language to written language [8]. Hori et al. used Weighted Finite State Transducers (WFSTs) for the same purpose by representing each component as a WFST [7]. Neibig et al. also used WFSTs, although their method utilized a WFST-based log-linear framework [9].

In this paper, we present a novel method for improving

the readability of ASR results based on a machine translation framework that uses a confusion network representing multiple hypotheses of the ASR results. Most methods for improving the readability of transcriptions have been developed for manual transcriptions and lack the ability of handling multiple hypotheses, thereby suffering severe degradation when dealing with ASR results [10]. By taking into account multiple hypotheses, the proposed method is expected to perform well under erroneous conditions in which the WER is relatively high. Multiple hypotheses of the ASR results are represented by a confusion network [11].

This paper is organized as follows. The formulation and details of our method are described in Section 2, while baselines used in the experiments are discussed in Section 3. The experimental setup and results are explained in Sections 4-6. Finally, Section 7 presents our conclusions and some future works.

## 2. Proposed Method

### 2.1. Formulation

The proposed method computes the probability of obtaining a written language style word sequence  $W$  from a sequence of speech  $O$  as follows:

$$\begin{aligned} P(W|O) &= \sum_S P(W, S|O) \\ &= \sum_S P(W|S, O)P(S|O) \\ &\approx \sum_S P(W|S)P(S|O), \end{aligned} \quad (1)$$

where  $S$  is a hidden variable and represents the word sequence of the spoken language.  $P(S|O)$  is the posterior probability of  $S$  given acoustic observation  $O$ , which can be derived from a confusion network.  $P(W|S)$  is the translation model between the written language  $W$  and spoken language  $S$ . we approximate  $P(W|S)$  as follows:

$$P(W|S) \doteq P(W)\delta(W, S), \quad (2)$$

where  $P(W)$  is a language model for the written language and  $\delta(W, S)$  is a function, the value of which is 1 if  $S$  can be converted into  $W$ , otherwise it is 0. For practical reasons, we compute  $\delta(W, S)$  ( $W = (w_1, w_2, \dots, w_N)$ ,  $S = (s_1, s_2, \dots, s_N)$ ) as follows:

$$\delta(W, S) = \prod_i^N \delta(w_i, s_i), \quad (3)$$

where  $\delta(w, s) = 1$  if the word pair exists in the translation table, otherwise 0. By using (2) and (3) and introducing weights to balance between  $P(W)$  and  $P(S_i|O)$ , (1) becomes

$$P(W|O) \approx P(W)^\alpha \sum_S \prod_i^N \delta(w_i, s_i) P(S_i|O)^\beta. \quad (4)$$

Our method represents multiple hypotheses (expressed as the summation of  $S$ ) using a confusion network and finds  $W$  that maximizes (4), while considering conversions from a spoken language network into a written language (represented as  $\prod_i^N \delta(w_i, s_i)$ ) and a written style language score (represented as  $P(W)$ ).

## 2.2. Conversion of Confusion Network

The confusion network is converted from a spoken language to a written language based on  $\prod_i^N \delta(w_i, s_i)$  in (4). By using (3), we can deal with any kind of translation. In this paper, however, we focus only on those aspects that should be dealt with first, i.e., the deletion of filled pauses, the insertion of commas and periods as punctuation marks, and the insertion of particles (preposition-like function words) which are often omitted in Japanese spontaneous speech [9].

### 2.2.1. Deletion of Filled Pauses

A filled pause is the most dominant of the specific phenomena in spontaneous speech [12], and has the greatest effect on the readability of the transcripts. To remove filled pauses, we added the following entries into the translation table:

$$\delta(\text{del}, \text{Filler}) = 1, \quad (5)$$

where *Filler* is a word whose part of speech (POS) is *filler* or *interjection*. It is difficult to remove filled pauses correctly from even the best transcripts, because such pauses are often mistaken for different words. However, our proposed method can detect and remove filler parts more accurately, since we sum the scores of the spoken words  $\{s|\delta(w, s) = 1\}$  that are converted into the same written word  $w$  based on (4). This means that if a bin contains many filled pauses, it tends to be deleted according to the majority decision principle.

### 2.2.2. Insertion of Punctuation Marks

The raw output from recognizers lacks punctuation marks and the unit used for recognition (usually a segment between short pauses) differs from the actual underlying sentence unit in the utterances. Therefore, to improve the readability of the output text, we need to detect sentence boundaries and to recover the punctuation marks. While punctuation marks do not always correspond to pauses, relationships do still exist between them [8]. Consequently, we added the following entries into the translation table:

$$\delta(\text{Punctuation}, \text{Pause}) = 1, \quad (6)$$

where *Punctuation* is a comma or period, and *Pause* is a silence or short pause. Furthermore, we also added the following entries into the translation table:

$$\delta(\text{Punctuation}, \text{del}) = 1. \quad (7)$$

By allowing a conversion from *del* to *Punctuation*, we can also recover punctuation marks that do not correspond to pauses.

### 2.2.3. Insertion of Particles

Since particles are often omitted in spoken Japanese, we need to estimate and recover these omitted particles. Such particles often do not correspond to any spoken words and are merely omissions between words. As a result, we consider particles only from the written *del*, and have added the following entries into the translation table:

$$\delta(\text{Particle}, \text{del}) = 1, \quad (8)$$

where *Particle* is a particle of “wa”, “ga”, “o”, and “to”, which are often omitted in Japanese [13].

## 2.3. Search by Stack Decoding

After the conversion of the confusion network from a spoken language style into a written language style, a search of the confusion network is needed to find the most plausible written style hypothesis  $\hat{W}$ . Thanks to the simplification of the translation probability  $P(W|S)$  by (2), we can perform the search process bin by bin and monotonously on the confusion network and can make use of bin-synchronous stack decoding for the beam, where sentence level hypotheses are retained in a stack up to the beam width.

In each bin, all hypotheses in the stack are popped, and each word in the current bin is connected to all the hypotheses popped from the stack. The new generated hypotheses are all rescored by taking account of the written language model score for the new connected word and the posterior of the connected word, in addition to the original score of the hypothesis. Thereafter, if hypotheses exist that share the same history within the range of the language model, these hypotheses are bundled together for efficient computation. Finally, the hypotheses with the highest scores are once again pushed onto the stack and are retained up to the beam width  $w$ . This process continues until the end of the confusion network, and the hypothesis at the top of the stack is selected as the final result.

## 3. Baselines

### 3.1. Filler Removal

As described in Section 2.2.1, filled pauses mostly affect the readability of the transcripts. Therefore, filled pauses are automatically removed from a transcript and the resulting transcript is regarded as the baseline. Filled pauses are removed from transcripts based solely on their POS information; thus, words with “filler” or “interjection” as their POS are removed. We refer to this baseline as *Filler-rm*.

### 3.2. Single Hypothesis Editing

Our proposed method uses multiple hypotheses (or a confusion network) to provide robustness against recognition errors. Consequently, it should be compared with a similar method except in that it uses a single hypothesis or a 1-best transcript instead of multiple hypotheses. Since the proposed method relies on a confusion network structure for editing, we need a confusion network that contains only a single best transcript for editing the 1-best transcript. For this purpose, we construct a confusion network from the 1-best transcript as depicted in Figure 1. Additionally, since we have to assign posteriors to each edge, we introduce a parameter  $c$  and assign  $c$  for *del* words and  $1 - c$  for all other words. By using the confusion network, we can edit the 1-best result using the proposed method. We refer to this baseline as *single-edit*.

## 4. Experimental Setup

We used 8 classroom lectures, each about 70 min long (about 12K words), from the CJLC corpus [12] as the test set. Manually transcribed texts (referred to as *manual* in the experimental results) are available for these lectures. In addition, we manually prepared transcriptions which were paraphrased from the spoken language style into a written language style (henceforth referred to as *paraphrased*) for each lecture (about 10k words).

We used the lectures in the CSJ corpus [14] to train 928 context dependent acoustic models of Japanese syllables. Each syllable model has a GMM to model the output distribution of the speech input feature vector, consisting of 12 MFCC,  $\Delta$  MFCC,  $\Delta\Delta$  MFCC,  $\Delta$  power and  $\Delta\Delta$  power. There are 4 mixtures in the GMM and the covariance matrix forms a block structure

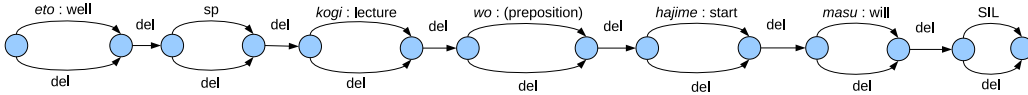


Figure 1: Construction of Confusion Network from 1-best result (E.g. *eto kogi wo hajime masu*: Well (I) will start (the) lecture).

Table 1: Recognition results of the test set [%].

Method	Del.	Ins.	Subs.	WER
MAP	7.3	4.8	32.4	44.4
ConfNet	11.0	3.0	28.6	42.5

Table 2: Word coverage and density of constructed confusion networks.

Input	Word coverage [%]	Word density
Manual	82.5	4.54
Paraphrased	77.2	

(full covariance for every MFCC,  $\Delta$  MFCC,  $\Delta\Delta$  MFCC,  $\Delta$  and  $\Delta\Delta$  of power).

Bi-grams with Witten-Bell discount are used to model the spoken language model for CJLC, while tri-grams with Witten-Bell discount are used to model the written language model. The spoken language model is trained on the CSJ corpus (2702 lectures), while the written language model is trained on the *Mainichi* newspaper corpus (9 years’ worth of articles). The vocabulary size is 20k and the lexicon is the set of words with higher frequencies in the CSJ corpus.

For decoding we used an in-house decoder called SPOJUS++, which has a feature to output a confusion network [15].

A lattice is pruned, leaving only 0.15% edges.  $\alpha$  and  $\beta$  in (4) are set to 1.0 and 6.0, respectively, while  $c$  in Section 3.2 is set to 0.05. The beam width  $w$  in Section 2.3 is set to 100.

## 5. Automatic Evaluation

### 5.1. Evaluation of Confusion Network

Table 1 gives the recognition results. In this table, MAP and ConfNet mean that the recognition results were computed using maximum a posteriori and consensus decoding [11], respectively. The consensus decoding improved the WER of the recognition results by 1.9%. We thus used the transcripts obtained by the consensus decoding as the 1-best results in the subsequent evaluations.

Table 2 shows the word coverage and density of the confusion networks, which contained 82.5% correct words for manual transcripts and 77.2% correct words for paraphrased transcripts. The word coverage for the paraphrased transcripts was smaller than that for the manual transcripts, since the transcriptionist was allowed to rewrite the expressions. The word coverage for the paraphrased transcripts was 77.2%, which was not enough to improve the understandability, but was sufficient to improve the readability as discussed in Section 6. The word density of the confusion networks was about 4.5.

### 5.2. Comparison with Paraphrased Transcripts

To evaluate the proposed method, we compared it (referred to in the results as *Proposed*) with different types of transcripts, namely, the raw transcript of the best recognition result (*1-best-raw*), the two baselines as described in Section 3 (*Filler-rm* and *Single-edit*), the raw manual transcript (*Manual-raw*), the manual transcript from which filled pauses were removed using the

Table 3: Evaluation results using the paraphrased transcripts [%].

Method	Particle	Del.	Ins.	Subs.	WER
1-best-raw	-	7.6	15.7	39.3	62.6
Filler-rm	-	10.0	9.7	36.5	56.2
Single-edit	yes	8.2	12.6	38.7	59.5
Single-edit	no	10.7	9.1	36.3	56.2
Proposed	yes	16.2	7.0	31.6	54.8
Proposed	no	18.9	5.2	29.6	<b>53.7</b>
Manual-raw	-	3.5	20.7	14.7	38.9
Manual-filler	-	4.0	14.4	13.7	32.1
Manual-edit	yes	5.8	16.0	17.1	38.9
Manual-edit	no	6.9	11.8	16.4	35.2

Table 4: Evaluation results for the insertion of periods.

Method	Recall	Precision	F
1-best-raw (pause)	0.562	0.253	0.335
Single-edit	0.660	0.278	0.384
Proposed	0.525	0.376	0.422
Manual-edit	0.634	0.394	0.465

method described in Section 3.1 (*Manual-filler*), and the transcript edited from the manual transcript using the method described in Section 3.2 (*Manual-edit*), which gives the upper-bound of the proposed method. Punctuation marks were removed from all transcripts for the evaluation. The results are given in Table 3. In the table, “particle” refers to whether (*yes*) or not (*no*) the insertion of particles described in Section 2.2.3 has been considered. The WER for the transcripts derived from the manual transcripts (*Manual-raw*, *Manual-filler*, *Manual-edit*) could not be 0% since our target was the paraphrased transcript.

From Table 3, it is clear that the insertion of particles did not provide any improvement, but instead degraded the performance in all cases, since the process caused a number of false insertions. (In order to reverse this degradation, we need more statistics on particles.) As a result, we did not consider the insertion of particles in the subsequent evaluations. The *Proposed* method outperformed both *Filler-rm* and *Single-edit*. This result shows that we can obtain further improvement by using multiple hypotheses. *Proposed* yielded 8.9% and 2.5% improvement against *1-best-raw* and *Single-edit*, respectively.

### 5.3. Evaluation of the Insertion of Periods

In this section, we evaluate the ability of the methods to insert periods. For this purpose, we aligned the transcripts with the paraphrased ones containing manually inserted periods and computed the *Recall*, *Precision* and *F-measure*.

Table 4 gives the evaluation results for the insertion of periods. Because the results for *1-best-raw*, *Filler-rm* and *Manual-filler* must be the same, the results given for *1-best-raw* are indicative of all three of these methods. The results show that *Proposed* outperformed *1-best-raw* and *Single-edit*. Using multiple hypotheses therefore also benefits the insertion of periods.

Table 5: Subjective test results.

Evaluation	Method		Count		
	A	B	A	B	?
Readability	1-best-raw	Filler-rm <sup>†</sup>	8	70	2
	Filler-rm	Proposed- <sup>†</sup>	23	51	6
	Proposed-	Proposed+ <sup>†</sup>	21	56	3
Understandability	1-best-raw	Filler-rm <sup>†</sup>	8	57	15
	Filler-rm	Proposed-	38	28	14
	Proposed-	Proposed+	24	38	18

## 6. Subjective Test

To assess whether the proposed method really improves the readability of the ASR results, we conducted a subjective test, with two lectures from the test set and 10 subjects, as follows:

1. Read transcripts A and B. Each transcript was about 30 lines (or 850 words) and was divided into four small blocks (about seven lines each), which roughly reflected the topics.
2. Compare each block for readability (paired comparison).
3. Read the manual transcript, and then reread/understand each transcript again.
4. Compare each block for understandability (paired comparison).

If there was no preference between a pair of transcripts, subjects were allowed to use “?”. Using the above procedure, we compared three different transcript pairs to evaluate each edited factor. In other words, we compared *1-best-raw* and *Filler-rm* to evaluate the effect of the removal of filled pauses; *Filler-rm* and *Proposed*, which used the same carriage return lines as *Filler-rm*, to evaluate the effect of the editing by the proposed method (*Proposed-*); and *Proposed-* and *Proposed*, which started new lines based on the inserted periods, to evaluate the effect of the punctuation (*Proposed+*). All punctuation marks were removed at the start to avoid any bias caused by them.

The results of the subjective test are shown in Table 5. In the table, “count” indicates how many times the method was chosen, while “†” denotes that the method achieved significantly better readability or understandability at the 1 significance level (z-test). The table shows that filler removal improved the readability and understandability significantly. This means that filled pauses should always be removed from transcripts. In contrast, *Proposed-* and *Proposed+* did not improve the understandability significantly, although they did significantly improve the readability. Because *Proposed* tries to remove disfluencies, thereby causing the deletion of content words that should be retained, the editing by the proposed method was not able to improve the understandability, although it did improve the readability. The results of the comparison of the carriage return line indicate that the accuracy of the insertion of periods shown in Table 4 was sufficient to improve the readability by splitting the transcript into a manageable size, but could not improve the understandability, because splitting at irrelevant points also reduces the coherence of the transcript. The subjective test, therefore, also confirmed the effectiveness of using multiple hypotheses.

## 7. Conclusion

In this paper, we presented a novel method for improving the readability of classroom lecture ASR results based on a machine translation framework that uses a confusion network. The proposed method constructs a confusion network to represent

multiple hypotheses and the most plausible paraphrased sentence is found using a written language model from the confusion network, which is converted from a spoken style into a written style by a translation model. The proposed method outperformed baselines in automatic evaluations, while the subjective test confirmed that the proposed method actually improved the readability.

In our future research, we need to take into account the correction of colloquial expressions, as in [8, 9], and the insertion of periods and particles using a discriminative model, which is difficult to implement in an unsupervised manner.

## 8. Acknowledgments

Part of this research was supported by Global COE Program “Frontiers of Intelligent Sensing” from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

## 9. References

- [1] Y. Fujii, K. Yamamoto, N. Kitaoka, and S. Nakagawa, “Class lecture summarization taking into account consecutiveness of important sentences,” in *Proc. Interspeech*, September 2008, pp. 2438–2441.
- [2] C. Chelba and A. Acero, “Indexing uncertainty for spoken document search,” in *Proc. Interspeech*, Sep. 2005, pp. 61–64.
- [3] S. Togashi and S. Nakagawa, “A browsing system for classroom lecture speech,” in *Proc. Interspeech*, Sep. 2008, pp. 2803–2806.
- [4] J. Glass, S. C. T. J. Hazen, I. Malioutov, D. Huynh, and R. Barzilay, “Recent progress in the MIT spoken lecture processing project,” in *Proc. Interspeech*, Aug. 2007, pp. 2553–2556.
- [5] S. Kogure, H. Nishizaki, M. Tsuchiya, K. Yamamoto, S. Togashi, and S. Nakagawa, “Speech recognition performance of CJLC: Corpus of Japanese lecture contents,” in *Proc. Interspeech*, Sep. 2008, pp. 1554–1557.
- [6] T. Kawahara, Y. Nemoto, and Y. Akita, “Automatic lecture transcription by exploiting presentation slide information for language model adaptation,” in *Proc. IEEE-ICASSP*, 2008, pp. 4929–4932.
- [7] T. Hori, D. Willet, and Y. Minami, “Paraphrasing spontaneous speech using weighted finite-state transducers,” *SSPR*, pp. 210–222, 4 2003.
- [8] K. Shitaoka, H. Nanjo, and T. Kawahara, “Automatic transformation of lecture transcription into document style using statistical framework,” in *Proc. Interspeech*, 2004, pp. 2881–2884.
- [9] G. Neubig, S. Mori, and T. Kawahara, “A WFST-based Log-linear Framework for Speaking-style Transformation,” in *Proc. Interspeech*, 2009, pp. 1495–1498.
- [10] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, September 2006.
- [11] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.
- [12] M. Tsuchiya, S. Kogure, H. Nishizaki, K. Ohta, and S. Nakagawa, “Developing Corpus of Japanese Classroom Lecture Speech Contents,” in *Proc. LREC*, 2008.
- [13] M. Yamamoto, s. Kobayashi, and S. Nakagawa, “An analysis and parsing method of the omission of post-position and inversion on Japanese spoken sentence in dialog (in Japanese),” *IPSJ*, vol. 33, no. 11, pp. 1322–1330, 11 1992.
- [14] S. Furui, K. Maekawa, and H. Isahara, “A Japanese national project on spontaneous speech corpus and processing technology,” *Proc. ASR2000*, pp. 244–248, 2000.
- [15] Y. Fujii, K. Yamamoto, and S. Nakagawa, “Improving the LVCSR System : SPOJUS++,” in *Proc of the Fourth Spoken Document Processing Workshop*, Feb. 2010, (in Japanese).