PAPER

Consolidation-Based Speech Translation and Evaluation Approach

Chiori HORI[†a)], ***Member***, **Bing ZHAO**[††], **Stephan VOGEL**[††], **Alex WAIBEL**[††], **Hideki KASHIOKA**[†], ***Nonmembers***, ***and*** **Satoshi NAKAMURA**[†], ***Member***

SUMMARY The performance of speech translation systems combining automatic speech recognition (ASR) and machine translation (MT) systems is degraded by redundant and irrelevant information caused by speaker disfluency and recognition errors. This paper proposes a new approach to translating speech recognition results through speech consolidation, which removes ASR errors and disfluencies and extracts meaningful phrases. A consolidation approach is spun off from speech summarization by word extraction from ASR 1-best. We extended the consolidation approach for confusion network (CN) and tested the performance using TED speech and confirmed the consolidation results preserved more meaningful phrases in comparison with the original ASR results. We applied the consolidation technique to speech translation. To test the performance of consolidation-based speech translation, Chinese broadcast news (BN) speech in RT04 were recognized, consolidated and then translated. The speech translation results via consolidation cannot be directly compared with gold standards in which all words in speech are translated because consolidation-based translations are partial translations. We would like to propose a new evaluation framework for partial translation by comparing them with the most similar set of words extracted from a word network created by merging *gradual summarizations* of the gold standard translation. The performance of consolidation-based MT results was evaluated using *BLEU*. We also propose *Information Preservation Accuracy* (IPAccy) and *Meaning Preservation Accuracy* (MPAccy) to evaluate consolidation and consolidation-based MT. We confirmed that consolidation contributed to the performance of speech translation.

key words: *speech translation, speech consolidation, TED speech, Chinese broadcast news speech, Chinese-English translation, BLEU, IPAccy, MPAccy*

1. Introduction

In the past, we have worked on spoken language processing system such as summarization of speech in meetings [1] and broadcast news [2] and machine translation (MT) of travel conversations in the C-star project [3], appointment negotiation in the Verbmobil project [4], and dialogue in e-commerce in the NESPOLE! Project [5]. The targets for these tasks are mainly restricted domains. The state-of-the-art speech recognition technology has been applicable to real world situations especially in terms of vocabulary size [6]. Now we are tackling domain unrestricted spoken language processing.

A spoken language processing system needs to be combined with language processing and automatic speech recognition (ASR) technologies. In the field of natural language processing, Summarization [7], Machine Translation (MT) [8], and Question Answering (QA) [9] on written text with large vocabulary such as newspaper text and HTML documents are being actively investigated using statistical approaches. Such technologies are incorporated into speech processing. However, written text is still difficult even if huge corpora are available for calculating statistic models and speech processing is even more complicated. The difficulty in speech processing is mainly caused by the style of spoken language being different from written text. Spontaneous speech includes colloquial expressions and ill-formed sentences caused by spontaneous aspects such as incorrect grammar, incomplete sentences, and redundant expressions i.e., disfluencies, repetitions, and word fragments. In addition, ASR output is not always perfect and we also have to handle recognition errors. There are mismatches between training data for statistical models used in language processing systems and ASR results of spontaneous speech.

Recently, spontaneous speech recognition has been intensively investigated. English academic presentation speech was recognized by adapting models of written text to spoken language transcriptions [10]. To detect phenomena in spoken language statistically, we need to collect spontaneous speech. Japanese academic presentation speech and free talk with various topics are manually transcribed and annotated precisely [11] and English broadcast news and conversational telephone speech are annotated with markers such as edit words in the EARS project [12]. Both ASR and language processing for domain unrestricted tasks have been actively researched. Now we are working on domain unrestricted speech translation tasks such as telephone conversations, lectures, meetings and broadcast news speech in the STR-DUST (Speech Translation for Domain-Unlimited Spontaneous Communication Tasks) project [13]. According to this research phase, we organized the International Workshop on Speech Summarization for Information Extraction and Machine Translation (IWSpS) [14], on spoken language processing including Summarization, MT and QA for domain unrestricted task. In this workshop, we attempted to solve degradation of total performance due to error-prone ASR results of spontaneous speech.

We have proposed speech consolidation as a process which cleans up speech transcription to enhance the total performance of spoken language processing. Speech con-

Manuscript received June 25, 2008.
Manuscript revised October 29, 2008.

[†]The authors are with NiCT-ATR, Kyoto-fu, 619–0288 Japan.
[††]The authors are with CMU, Carnegie Mellon University, School of Computer Science, interACT, 407 S. Craig Street Pittsburgh, PA 15213, USA.
a) E-mail: chiori.hori@nict.jp

DOI: 10.1587/transinf.E92.D.477

solidation removes redundant information caused by disfluencies and irrelevant information by recognition errors from ASR results to alleviate mismatches between statistical models for spoken language processing and spontaneous speech. A straightforward strategy to consolidate speech recognition is to remove disfluent and unreliable phrases from ASR results. To handle disfluency, such as fillers, repetitions, corrections and false starts, a disfluency removal approach has been proposed [15]. Additionally we also have to handle recognition errors. Focusing on deleting misrecognised Out-Of-Vocabulary (OOV), OOV words are forcibly recognized as a word in the ASR vocabulary and so not only the OOV word itself but the words surrounded it are accidentally misrecognized. Detecting OOV words is generally difficult in domain unrestricted tasks. On the other hand, confidence measures [16] can be applied to delete acoustically and linguistically unreliable phrases. However, the meaning of consolidation results made by just removing unreliable phrases from ASR results based on the approaches described above sometimes do not correspond to the original meaning intended by the speaker. Consolidation should preserve original meanings.

We have proposed a statistical summarization approach which extracts words from transcriptions according to compression ratios by focusing on 1) extracting important content words, 2) excluding redundant and irrelevant phrases, and 3) concatenating words in this summarization approach to maintain original meanings [2]. This approach accomplishes important information extraction and speech consolidation simultaneously. The preliminary experiments of summarization-based translation shows that extraction of more reliable phrases which preserve the original meaning could contribute positively to the total performance of MT. In this paper, we extend consolidation aspects in summarization by combining functions 2) and 3). The consolidation approach proposed here attempts to extract meaningful phrases which maintain the original meanings as long as possible without being given a compression ratio.

This paper reports preliminary experiment results for a summarization-based speech translation in which English news speech was translated into Chinese text in Sect. 2. A speech consolidation approach is described in Sect. 3. To evaluate the performance of consolidation, we propose an evaluation framework using *gradual summarization networks* based on *Information Preservation Accuracy* (*IPAccy*) and *Meaning Preservation Accuracy* (*MPAccy*) [19] in Sect. 4. Evaluation results of automatic speech consolidation on lecture speech in the international conference recorded in the Translanguage English Database (TED) corpus [17] are described in Sect. 5.

Furthermore, this paper presents a consolidation-based speech translation system in which ASR [18], consolidation [19] and statistical MT [20], [21] systems are cascaded. Mandarin news speech in RT04 was recognized, consolidated, and translated into English text. We evaluated both intrinsic and extrinsic performance of the speech consolidation i.e., consolidation accuracy and MT performance effected by consolidation, respectively. When evaluating the consolidation-based speech translation, the speech translation results via consolidation cannot be directly compared with the gold standard in which all words in speech are translated. This is because consolidation-based translations are partial translations and do not always preserve all information in original speech. To evaluate such partial translations properly, this paper proposes an evaluation framework by comparing with the most similar set of words extracted from a word network generated by merging gradual summarizations of the gold standard translation based on *Meaning Preservation Accuracy* (*MPAccy*) [22]. The evaluation results for consolidation of Chinese BN is reported in Sect. 5. The evaluation framework and the evaluation results of consolidation-based translation is presented in Sects. 7 and 8, respectively.

2. Summarization-Based Speech Translation

A summarization approach extracts words from ASR output according to a given compression ratio to maximize a summarization score using a Dynamic Programming (DP) technique. The summarization score indicating the appropriateness of a summarized sentence is defined as the sum of the linguistic score L of the word string in the summarized sentence, the word significance score I, the confidence score C of each word in the original sentence and the word concatenation score Tr. The word concatenation score given by SDCFG (Stochastic Dependency Context Free Grammar) indicates a word concatenation probability determined by a dependency structure in an original sentence.

Given a ASR result consisting of N words, $W = w_1, w_2, \ldots, w_N$, the summarization is performed by extracting a set of $M (M < N)$ words, $V = v_1, v_2, \ldots, v_M$, which maximizes the score given by Eq. (1).

$$S(V) = \sum_{m=1}^{M} \lambda_L L(v_m | v_1 \ldots v_{m-1}) + \lambda_I I(v_m) + \lambda_C C(v_m) + \lambda_T Tr(v_{m-1}, v_m), \quad (1)$$

where λ_L, λ_I, λ_C and λ_T are weighting factors for balancing among L, I, C, and Tr.

We tested whether the consolidation function in the summarization could contribute to the quality of machine translation in the IWSpS [14]. Five news stories, consisting of 25 utterances on average, of CNN broadcast news speech were transcribed and summarized at 40% and 70% extraction ratio [2]. The word accuracy of the ASR results was 78.4%. The summarization results were translated into Chinese using our statistical machine translation [20]. The translation results were evaluated by one human subject in terms of linguistically correctness (*readability*), ratio of information preservation (*extraction*), correctness of preserved information (*correnctness*) and quality in total (*whole*). A Chinese native speaker read 250 Chinese sentences translated from the transcription of English news speech and ranked them from 1 to 5, with 5 being the best

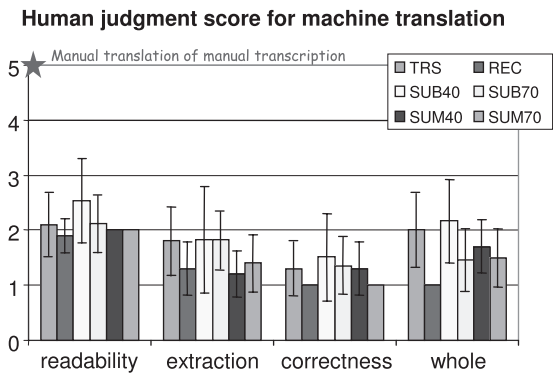

Fig. 1 Human judgment for machine translation of speech summarization results. TRS: manual transcription, REC: ASR results, SUB40 and SUB70: manual summarization at 40% and 70% ratio, SUM40 and SUM70: automatic summarization at 40% and 70%.

score.

Figure 1 shows the results of the human judgment. Manual translation of manual transcription is a gold standard and given a score "5" as best translations. TRS and REC represent translation of manual transcription and ASR results respectively. SUB and SUM represent manual summarization and automatic summarization respectively. The suffix numbers, i.e., 40 and 70, are summarization ratios.

The quality of 40% manual summarization is given the best score in all conditions among 6 groups, while the quality of the full translation using ASR results is always given the lowest score. These results indicate removing disfluent and unreliable phrase, which preserving original meaning has the potential to enhance the performance of MT.

Furthermore, we found another benefit in *summarization based translation*. The state-of-the-art technology of statistical MT does not work well especially when translating language pairs which have totally different syntactic structures. The current word reordering model is not enough to handle this problem. In our results, the quality of 40% summarization is much better than that of 70% in all evaluation points. This is because English-Chinese translation has a reordering problem and longer sentences are liable to be translated worse than shorter ones.

3. Speech Consolidation

3.1 Consolidation Approach

In consolidation, removing recognition errors, retaining as much information of the original sentence as possible and reconstructing a fluent sentence are important factors. There is no restriction by giving a compression ratio like summarization. We modify the summarization score to function for effective consolidation as

$$S(V) = \sum_{m=1}^{M} \{\lambda_L L(v_m|v_1 \ldots v_{m-1}) + \lambda_C C(v_m) + sp \cdot d(v_{m-1}, v_m) + ip\}, \quad (2)$$

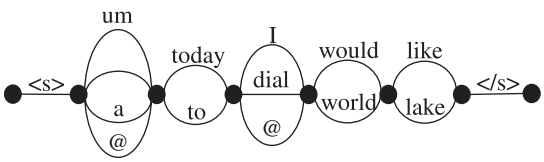

Fig. 2 An example of confusion network.

where sp is a skip penalty ($sp < 0$); $d(v_{m-1}, v_m)$ is the number of skipped words between v_{m-1} and v_m; ip is a insertion penalty. We proposed a dependency score to avoid concatenating two words which do not have a dependency structure [2]. However, dependency detection of spontaneous speech is still challenging. We used the skip penalty instead of the dependency score in this study. The sp is incorporated to avoid connecting two words located a long distance apart in the original sentence. While the insertion penalty (ip) is incorporated to avoid high compression of the original sentence (i.e. low summarization ratio) because high compression of a sentence often alters the meaning of the sentence.

The linguistic score $L(v_m|v_1, \ldots, v_{m-1})$ indicates the appropriateness of the word strings in a summarized sentence. It is measured by the logarithmic value of a trigram probability $P(v_m|v_{m-2}, v_{m-1})$. For consolidation, since we focus only on connectivity between words, we use an adjusted trigram probability $P(v_m|v_{m-2}, v_{m-1})/P(v_m)$ instead of the regular trigram. This normalized trigram removes the influence of frequency and represents only word concatenation correctness.

The confidence score $C(v_m)$ is incorporated in the above equation to weight acoustically as well as linguistically reliable hypotheses. Specifically, a posterior probability of each transcribed word, that is the ratio of a word hypothesis probability to that of all other hypotheses, is calculated using a word graph obtained by a decoder and used as the confidence measure. In this study, we use a confusion network [24] instead of a word graph since more accurate posterior probabilities are derived from confusion networks.

3.2 Consolidation of Confusion Network

A DP technique for speech summarization can directly be applied to speech consolidation. However, the algorithm is only for a 1-best hypothesis in speech recognition. In this work, we extended the algorithm to find the best consolidation result from among multiple hypotheses represented in a confusion network [24]. The extended algorithm has the potential to reduce recognition errors by reselecting the words in the network through the consolidation.

A confusion network is a compact representation of multiple hypotheses generated in speech recognition. Figure 2 shows an example of the confusion network. Compared to a word lattice/graph, it is more compact since it ignores the connectivity of adjacent words and discards time information of each word. We assume that all sentences included in a confusion network begin with "<s>" and end with "</s>". Let N be the length of the confusion network,

Table 1 An example of a manual consolidation.

REF	which is another topic of interest such as in identifying the sex or the language sex of the speaker the identity of the speaker or the language being spoken
CON	which is another topic of interest such as in identifying the sex [***for*** the language sex of speaker], [***given state*** speaker ***of***]the language being spoken

REF: manual transcription and CON: manually consolidated ASR output, bold and italic words indicate recognition errors and phrases bracketed are removed.

i.e. the number of confusion sets. The confusion set consists of a set of competing words in one column as in Fig. 2.

For example, the first confusion set includes only "<s>", and the second set includes "um", "ah", "a", and "@". The symbol "@" is a special word indicating a possibility of deletion. In a confusion network, a posterior probability is attached to every word. The sum of the probabilities in each confusion set becomes 1. The algorithm of confusion network based consolidation is described in Appendix A.

4. Evaluation Framework for Consolidation

4.1 Gold Standard Based on Manual Consolidation

The most ideal gold standards for each ASR results are prepared by humans. Table 1 shows an example of manual consolidation. 1-best of ASR output was manually consolidated by deleting disfluent expressions and phrases which have the different meaning from the manual transcription. Since recognition errors can be changed under different recognition conditions, we need gold standards for each recognition result. However, such gold standards prepared for each recognition result are very expensive. To alleviate this labor, this paper proposes an evaluation method using *gradual summarization networks*.

4.2 Gold Standard Based on a Gradual Summarization Network

The goal of consolidation is to extract meaningful phrases which preserve part of the original meaning from ASR outputs by removing recognition errors and meaningless fragments. To generate gold standards for this goal, we can delete substitution and insertion errors from ASR results by comparing with manual transcriptions at the beginning. Supposing an utterance (TRS) is misrecognized as ASR in Table 2, we can delete recognition errors i.e., "boot" and "chill" from TRS. We denote the word string excluding recognition errors as DEL.

Although the DEL consists of only words which are recognized correctly, some fragments, i.e. "The", "in" still remain in the DEL. To obtain a gold standard for consolidation, we also need to delete such meaningless fragments

Table 2 Example of ASR and consolidation results.

TRS	<s> The beautiful cherry blossoms in Japan bloom in spring </s>
ASR	<s> The (boot) (chill) blossoms in Japan bloom in _____ </s>
DEL	<s> The_____ blossoms in Japan bloom in _____ </s>
CON	<s>_____ blossoms in Japan bloom __ _____ </s>

TRS: transcription, ASR: recognition results,
DEL: ASR results excluding errors, CON: consolidation results

Table 3 Example of gradual summarization.

ORG	<s> The beautiful cherry blossoms in Japan bloom in spring </s>
GRD	<s> The _____ cherry blossoms in Japan bloom in spring </s>
	<s> The _____ cherry blossoms __ ____ bloom in spring </s>
	<s> The_____ cherry blossoms __ ____ _____ in spring </s>
	<s> The _____ cherry blossoms __ ____ _____ __ _____ </s>
	<s> ___ _____ _____ blossoms __ ____ _____ __ _____ </s>

ORG: manual transcription, GRD: gradual summarization of ORG

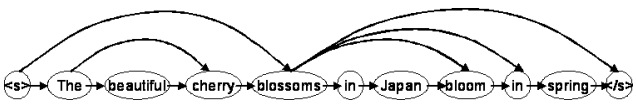

Fig. 3 Example of a gradual summarization network.

from the DEL. Humans can delete such fragments by considering the context, and get the ideal consolidation result (CON), i.e. the gold standard. However, since human labor for consolidating each ASR result is too expensive, we need to clean up such fragments automatically. When dependency structure of the TRS is known, it is not difficult to detect fragments. However the dependency structure of spoken language is still difficult to estimate accurately. In this evaluation, we use *gradual summarization networks*.

The gradual summarization is done manually by removing modifiers gradually to retain the original meaning of TRS as much as they can. Since significance is also considered, heads are sometimes removed earlier before removing modifiers. The gradual summarizations and the original sentence are then merged into a word network. We assume that the summarization network represents a set of sentences which maintain the original meaning because this network contains only word concatenations appeared in the merged strings and does not allow choosing meaningless fragments. When we evaluate a consolidated result, we first generate DEL, i.e. remove the recognition errors from the consolidated sentence, and choose the most similar word string to the DEL from the network as a gold standard.

Suppose the original utterance (TRS) in Table 2 is manually summarized by deleting words step by step as shown in Table 3. These gradual summarizations can be merged into a word network in Fig. 3. We consider all word strings

from <s> to </s> in the network retain part of the original meaning.

The gold standard can be selected based on Levenshtein distance from the network. For the recognition result in Table 2, "blossoms in Japan bloom" is selected as the gold standard, where all fragments as appeared in DEL are not included. Once gradual summarizations for each transcription are made, we can prepare the gold standard for each ASR result automatically using the gradual summarization network.

We note that the gradual summarization network does not cover all word strings which retain part of the original meaning. However, these missing word strings which preserve the original meaning do not favor automatic consolidation.

4.3 Evaluation Metric

To evaluate the performance of automatic consolidation, we compare the consolidation results with the gold standard which is extracted from gradual summarization network in terms of similarity. We define two types of measures based on word accuracy. One is for meaning preservation and the other is for information preservation. We denote *meaning preservation accuracy* and information preservation accuracy as *MPAccy* and *IPAccy*, respectively.

MPAccy shows to what degree the consolidation results preserve the original meaning. The degree of preservation of the original meaning via consolidation is evaluated by comparing a consolidation result and an extracted gold standard from a gradual summarization network.

IPAccy shows how much information in the original utterance is preserved. The compression ratio of the gold standards shows the upper bound of information preservation by consolidation. Since *MPAccy* shows what degree of information in the gold standard can be preserved by consolidation, we can see the total performance against the original utterance using Eq. (3).

$$IPAccy = MPAccy * \mathrm{CR}(\text{Gold Standard}) \quad (3)$$

where CR shows the compression ratio of the extracted gold standard. Our consolidation approach cannot preserve the same amount of information in original speech due to recognition errors and thus *IPAccy* has an upper bound. On the other hand, the consolidation has a potential to achieve 100% *MPAccy*.

5. Evaluation Experiments for Consolidation

5.1 Consolidation of Speech in the TED

English academic presentation speech in the TED corpus automatically transcribed using the Janus Recognition Toolkit (JRTk) in the IWSpS was used for evaluation experiments. Eight talks were recognized and evaluated by comparing manual consolidations by human.

5.1.1 Experimental Condition

ASR system

Eight talks were recognized with an acoustic model trained on 300 hours of Broadcast News (BN) data merged with the close talking channel of meeting corpora. The acoustic model used 42 features and consisted of 300 k Gaussians with diagonal covariances organized in 24 k distributions over 6 k codebooks [18]. The language model (LM) used for the speech recognizer was generated by interpolating a word 3-gram and a class-based 5-gram LM each trained on BN data (160 M words) and the proceedings corpus (see Sect. 2.1.3), and a 3-gram LM based on talks (60 k words) by the TED adaptation speakers. The overall OOV rate is 0.3% with a vocabulary size of 25000 words including multi-words and pronunciation variants. The average word error rate of the talks used in this paper is 33.3%.

Consolidation module

Linguistic score was calculated using BN data (160 M words) and the proceedings corpus (17 M words). Confidence score obtained from a confusion network by the ASR system was applied. We separated the eight talks into two sets, one is used as a development set and the other is used as a test set, each of which consists of four talks. The best scaling factors for consolidation scores were experimentally determined using the development set. The test set was evaluated based on the best scaling factors.

Gold standard for consolidation

1-best of ASR output was manually consolidated by deleting disfluent expressions and phrases which have different meanings from the manual transcription by a human.

5.1.2 Evaluation Results

First, we investigate the accuracy of automatic consolidation, and the effectiveness of each score used in the consolidation algorithm. Figure 4 shows the word accuracy of consolidation results derived from 1-best ASR output in the development set and the test set. L, C, and sp indicate the use of linguistic score, confidence score, and skip penalty, respectively. For example, "L+sp" shows the case where

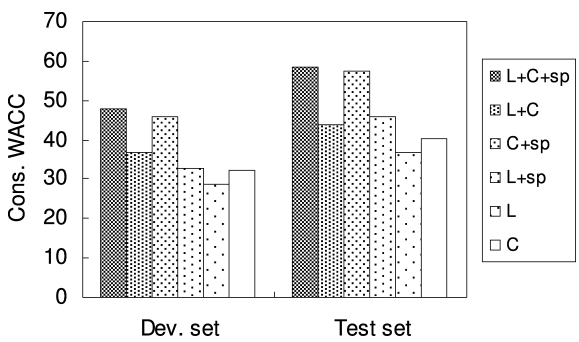

Fig. 4 Word accuracy of consolidation results.

Table 4 Ratio of extracted words in spoken words and ratio of correctly recognized words in consolidation results.

LectureID	Manual/1-best		Auto/1-best		Auto/ConfNet		ASR
(TestSet)	Ratio%	Prec%	Ratio%	Prec%	Ratio%	Prec%	WACC
dc57s200	64.2	100.0	71.3	88.0	70.6	88.1	67.2
hb64s400	48.1	100.0	62.9	87.7	62.3	87.9	66.6
ro31s400	79.5	100.0	72.7	95.2	72.1	95.2	83.3
yi59s500	57.9	100.0	74.2	86.7	73.4	86.8	65.7
total	66.6	100.0	71.1	90.5	70.4	90.6	72.5

only a linguistic score and a skip penalty are used for consolidation. In both sets, "L+C+sp" gave the best accuracy. Hence all scores defined in this paper are effective for consolidation. Although confidence score seems to be dominant compared to the other scores, a high accuracy was not derived using only the confidence score.

Next, we will discuss properties of the consolidation results. Table 4 shows the ratio of the number of automatically extracted words in consolidation to the number of spoken words (Ratio%) for "L+C+sp". We also calculated the ratio of the correctly recognized words contained in the consolidation results, i.e. the precision (Prec%) of in evaluating the performance of removing recognition errors. We compared three cases of manual consolidation from 1-best ASR output (Manual/1-best), automatic consolidation from 1-best ASR output (Auto/1-best), automatic consolidation from confusion networks (Auto/ ConfNet). For reference, word accuracy of 1-best hypothesis in speech recognition (ASR WACC) is also attached in the table.

In Manual/1-best, it is shown that the human subject extracted 66.6% of spoken words in total. Since the subject knew which words were misrecognized, the precision resulted in 100%. On the other hand, as shown in Auto/1-best, the consolidation module selected 71.1% of spoken words, which was a similar value to that of Manual/1-best. In each lecture, however, the ratio was not so similar. Although the human subject tends to extract more words from the ASR output with higher word accuracy, such a tendency did not appear in the results of automatic consolidation.

In this experiment, we could not confirm the efficiency of applying consolidation to confusion networks since the result of Auto/ConfNet is almost the same as that of Auto/1-best. However, in both cases, it is shown that the consolidation method can extract accurately recognized words with high precision of above 90%.

5.2 Consolidation of Chinese BN Speech

We applied the consolidation approach to Chinese BN speech in RT04. To test the performance of consolidation, we used 125 utterances extracted from the beginning of 297 utterances in RT04 which were recognized and consolidated. The manual transcription for all speech are provided by LDC[†].

5.2.1 Experimental Condition

Test set of RT04 data involves English speech within Mandarin speech and thus some speakers were entirely misrecognized. In addition, some speech data are dialogues with a very spontaneous style.

ASR system

The ISL RT04 Mandarin Broadcast News evaluation system using the JANUS speech recognition toolkit was applied to the speech translation system [2]. The acoustic models were trained using 27 hours of the Mandarin HUB4 1997 training set and 69 hours of the TDT4 Mandarin data. 42-dimension features after Linear Discriminant Analysis were used for the front-end processing. The system employs a multi-pass decoding strategy in which cross adaptation among the syllable-based and the phone-based decoders were performed.

We used several corpora for our language model (LM) development: Mandarin Chinese News Text (LDC95T13), TDT{2,3,4}, Xinhua News, People's Daily and China Radio respectively are contained in the Mandarin Gigaword corpus and the HUB4 1997 acoustic training transcript. The vocabulary size is 63 K words. Confusion word networks were given to the consolidation system.

The character and word errors of our ASR system were 21.2% and 46.8%, respectively. The word accuracy was affected by mismatches of word segmentation between ASR output and manual transcription.

Consolidation module

Language model used in the ASR system was applied to consolidation. Since the performance of consolidation using 1-best ASR results was almost the same as that using confusion network, we consolidated 1-best ASR results for this task.

Gold standard for consolidation

To generate gold standards for consolidation, a human gradually summarizes manual transcription by deleting words as described in Sect. 4.1. The most similar word string was extracted from a gradual summarization network as a gold standard for each ASR result.

5.2.2 Evaluation Result

Figure 5 shows the results of the intrinsic evaluation based on character. The compression ratio of the consolidation is almost the same as the gold standard. The consolidation approach worked well to detect the length automatically.

IPAccy indicates that consolidation preserved more

[†]http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005S16

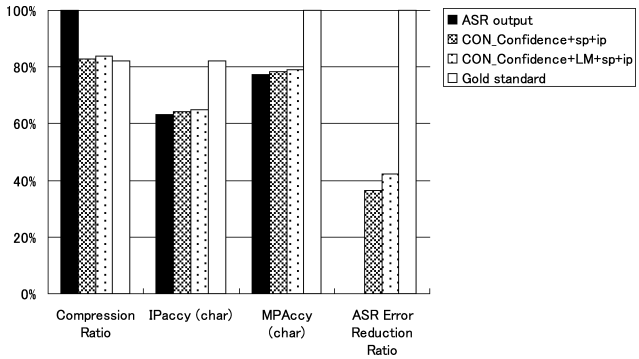

Fig. 5 Performance of consolidation of Chinese BN based on character.

Fig. 6 Consolidation-based speech translation system.

original information than the ASR results did. *MPAccy* of the ASR results was 77.2%; our consolidation approach achieved 79.2% *MPAccy*. To evaluate how many recognition errors are removed from ASR results, the reduction ratio of speech recognition errors were calculated. 42.4% of insertion and substitution errors in the ASR output were removed by consolidation. Our consolidation achieved a higher *MPAccy* and a higher reduction ratio of recognition errors than those for the ASR output.

These results show that our consolidation approach can extract a set of phrases which preserve the original meaning by excluding fragments and recognition errors.

6. Consolidation Based Speech Translation

To alleviate the degradation of the performance of speech translation, we proposed a new approach to translate ASR results through consolidation. To test whether the consolidation function can contribute to the quality of machine translation or not, we translated the consolidation results of Chinese BN speech to English text. Our system is designed as ASR [18], speech consolidation (SPCON) [19] and statistical MT (SMT) systems [20], [21] are cascaded in Fig. 6.

7. Evaluation Framework for Consolidation-Based Speech Translation

The whole cascaded system of ASR, consolidation and MT systems is evaluated in terms of how much consolidation can enhance the performance of speech translation. Since consolidation results missed some phrases in speech, we cannot evaluate the real contribution by consolidation when comparing with references in which all words in the source speech are translated. To evaluate partial translation based on consolidation, we need manual translations for each partial transcription.

Table 5 Example of synchronous gradual summarization.

	Source language	Target language
	Transcription	Translation
Original	A B C D E F	d f e c b a
Step 1	A B D F	d f b a
Step 2	A D F	d f a
Step 3	A F	f a

7.1 Gold Standard Based on Synchronous Gradual Summarization

The ideal gold standards for consolidation-based speech translation are manual translations of all "part of phrases" which retain part of the original meaning in speech. To obtain translation of "part of phrases", the bilingual translator generates one manual translation in response to each utterance in the source side and then gradually summarizes both the source and target sides by extracting words synchronously. Table 5 shows an example of a process to generate *synchronous gradual summarizations*.

A set of words, "A B C D E F", in the source side is translated into a set of words, "d f e c b a", in the target side in this example. The transcription and its translation are gradually summarized according to 3 steps. Each set of gradual summarizations of both sides has the same meaning. We received multiple manual translations of various lengths. To cover more word strings which preserve part of the original meaning in the manual translations, the gradual summarizations of the manual translations were merged into a word network as described in Sect. 4.1. A set of words minimizing errors based on the Levenshtein distance is extracted for each consolidation-based translation. We use the extracted string as a gold standard for consolidation and evaluated the similarity between it and the consolidation-based MT.

7.2 Evaluation Metric

The performance of the consolidation-based MT results were evaluated using the BLEU and *MPAccy* described in Sect. 4.3.

8. Evaluation Experiment for Consolidation Based MT

8.1 Experimental Condition

Test set is RT04 consisting of 297 utterances segmented for evaluation. We translated the consolidation results reported in Sect. 5.2.

Synchronous gradual summarization

To generate gold standards for consolidation and consolidation-based MT, the bilingual translator translated the 125 Chinese manual transcriptions into English text and then gradually summarized both the Chinese and English sides by extracting words synchronously. The Chinese gradual summarization was used in the intrinsic evaluation for

Table 6 Data used for phrase alignment model.

	#sentences	#words	#characters
Trains04.split.en	760471	12524365	72357758
Train04.split.gb	760471	11484629	46091860

the consolidation accuracy in Sect. 5 and the English gradual summarization was used in the extrinsic evaluation for the consolidation-based translation.

Gold standard for consolidation-based translation

We prepare three types of references for MT, i.e., one reference which is a manual translation for each manual transcription and multiple references which are manual translations for all gradual summarizations. To prepare references for consolidation-based MT, the gradual summarization in the target side are merged into a word network. A set of words maximizing word accuracy is extracted from the *gradual summarization network* by comparing it with consolidation-based MT. The length of the extracted word string is almost the same as that of the consolidation-based MT. We set the extracted word string as a certain reference.

Statistical Machine Translation system

Manual transcription of speech and recognition results were translated using the CMU SMT system based on phrase-to-phrase translations [20]. The experimental conditions were as follows:

Phrase alignment model:

760,471 sentence pairs were sub-sampled for the TIDES '02, '03 and '04 test sets from a 200 million words parallel corpus. The feature of data is listed in Table 6. Phrase table contains 1,666,428 entries ranging from 1-gram to 10-gram on the source side. There are eight score functions for each phrase pair [21].

Baseline performance using Chinese newspaper text:

The SMT system was constructed for translating Chinese newspaper text. The test sets provided in TIDES '02 can be translated with BLEU=27.22 (length penalty=1) and NIST= 8.7143 (length penalty=0.9942).

MT Performance for Manual transcription:

We split 297 utterances into 627 sentences based on full stops inserted by a human, which are different from the ASR 1-best segmentations. The final best scores we obtained is BLEU=9.85 (length penalty=0.995, Ratio of target against source=0.995). We also translated original utterance including multiple sentences. The scores were BLEU=8.59 and NIST=4.2730. These scores were slightly lower than those of segmentation based on sentences.

The evaluation results show the poor MT performance. The approach is not sufficient to accomplish the correct translation. In addition, it is difficult to cover all gold standards for the MT. There remains a large difference between manual translation and MT, which is mainly based on word-to-word translation, although our translator tried to translate word-to-word.

Even when each utterance was split into ideal sentences, the performance for translating BN was drastically degraded in comparison with translating a newspaper text.

Table 7 Out-of-vocabulary rate and perplexity.

			#sentences	#words	OOV	PP
TIDES '02	Source		878	24337	0	138
	4 references		3512	105143	0	148
RT04	Source	TRS	297	11547	0	536
		ASR	297	9724	0	848
	1 reference		297	12105	0	300

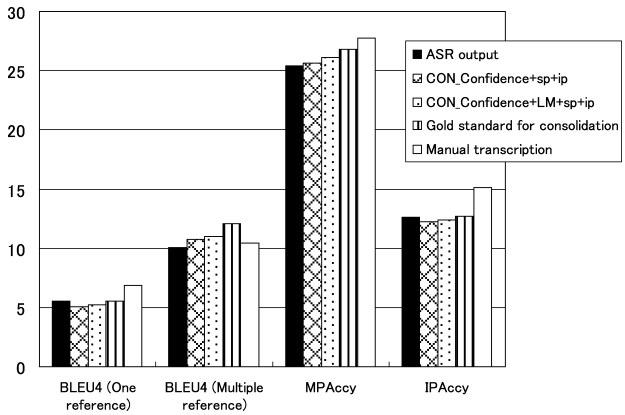

Fig. 7 Performance of machine translation.

The degradation was caused by model mismatch between training data and test data.

MT Performance for ASR output:

Speech recognition output with 21.2 % character error rate was translated with BLEU=8.20 and NIST=4.1425. The difference between translating the manual transcription and the ASR output is not significant. The results show that the degradation of the performance of translating BN against translating newspaper text is mainly caused by mismatch features in the model between BN and newspaper text. Table 7 lists out-of-vocabulary (OOV) rate and perplexity (PP) for each test sets.

The OOV and PP for the source side is calculated using a trigram of the source text in the parallel corpus for the alignment model and those for the target side is calculated using a trigram used in the SMT decoder. The vocabulary size of the source side (Chinese) is 107,829 and that of the target side (English) is 104,351.

8.2 Evaluation Result

Figure 7 shows the extrinsic evaluation for consolidation-based MT. The MT of the manual transcription, the ASR results, and the consolidation with and without language model score were evaluated. The lengths of the MT of the manual transcription, the ASR output, and the consolidation result were 74%, 67%, and 60% of the manual translation of the manual transcription, respectively.

We evaluated the MT based on BLEU (N=4) using one reference. There was no difference among the MT with or without consolidation. The BLEU using all gradual summarizations with different lengths shows that the consolidation contributed to enhance the MT performance. However,

that of the MT of the manual transcription was not evaluated well. *MPAccy* using a certain reference for the MT extracted from the gradual summarization network shows the MT performance was enhanced by the consolidation.

8.3 Discussion

Figure 6 in Sect. 5.2 shows that the consolidation preserved part of the meanings of the original speech excluding the errors, but the BLEU using one reference does not show a difference in the MT performance between the ASR output and the gold standard for consolidation in Fig. 7. This is because the partial translations are not counted as a correct translation even if the partial translation preserves part of the original meaning. When we used each gradual summarizations as multiple references, the BLEU was increased by the consolidation. However, the BLEU of the manual transcription was lower than those of others. There is a problem in the evaluation based on BLEU using multiple references with various lengths. The precision-based BLEU for shorter translations tend to be higher than that of longer translations. To give a penalty to precision for shorter translations, BLEU is penalized by length of references when hypotheses are shorter than references. When we considered multiple references with various lengths, all translations can find references which have the same length or similar length and thus the penalty for length given for BLEU does not work well. The results just show that the "longer" translations resulted in "lower" precision. On the other hand, the translations of the manual transcriptions were fairly evaluated by *MPAccy* using the gradual summarization network. We confirmed that our proposed method (CON_confidence +LM+sp+LM) outperformed the direct translation of ASR result (ASR output) in MPAccy score, where the significance level was 0.05. To test the statistical significance, we assumed that the upper bound of MPAccy score was that of the translation result for manual transcriptions since the MPAccy score of consolidation-based translation is not principally beyond the upper bound. The results of *MPAccy* show that the consolidation enhanced the performance of MT, while the *IPAccy* of the consolidation-based translation is almost the same as that of the ASR results. This means consolidation can preserve more meaningful phrases even some information is missed.

9. Conclusion

This paper proposes a new approach to speech translation through speech consolidation. To evaluation consolidation and consolidation-based speech translation, this paper proposes an evaluation framework using gradual summarization networks and evaluation metrics i.e., *IPAccy* and *MPAccy*.

TED speech was recognized and consolidated. The RT04 Mandarin Broadcast News speech was recognized, consolidated and translated into English text. We confirmed that our consolidation approach can extract a set of phrases which preserve the original meaning by excluding fragments and recognition errors. Furthermore, MT performance is enhanced by the consolidation. *MPAccy* using multiple references consisting of gradual summarizations of manual translation is capable of evaluating consolidation-based MT reasonably. We used consolidation of 1-best ASR results for speech translation for Chinese BN because there is no significant difference between 1-best and confusion network translation for TED speech.

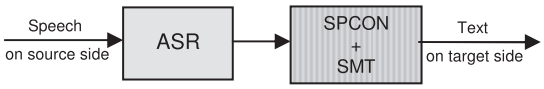

Fig. 8 Consolidation-based speech translation on confusion network.

Recently, speech translation is done on confusion networks (CN) obtained by ASR systems [25] since CN is compact and capable to keep multiple hypotheses [24]. Word reordering in MT using a CN is much easier than that using an ASR word lattice. Since the performance of CN consolidation is comparable with 1-best consolidation, we can use CN consolidation and CN translation without degradation from translation using 1-best consolidation. We can integrate CN consolidation directly into MT systems that translate from CN. This approach has a potential to select more reliable and meaningful phrases on the source side and reordering more combinations of word sets on the target side simultaneously. The hierarchical integration of ASR, SPCON and SMT shown in Fig. 6 can be modified as shown in Fig. 8.

This integration can be done as follows:

$$
\begin{aligned}
\hat{T} &= \arg\max_{T} P(T|O) \\
&= \arg\max_{T} \sum_{S_{CON}} \sum_{W} P(T|S_{CON}, W, O)\, P(S_{CON}|W, O) \\
&\quad \cdot P(W|O) \\
&\approx \arg\max_{T} \sum_{S_{CON}} \sum_{W} P(T|S_{CON})\, P(S_{CON}|W)\, P(W|O) \\
&\approx \arg\max_{T} \left[\max_{S_{CON}, W} P(T)\, P(S_{CON}|T)\, P(S_{CON}|W) \right. \\
&\quad \left. \cdot P(W)\, P(O|W) \right]
\end{aligned}
\tag{4}
$$

O : Speech input of source language (observed)
W : ASR result of the source language
S_{con} : Consolidated source language
T : Target Language (translation result)
$P(O|W)$: Acoustic model (speech recognition)
$P(W)$: Language model (source language)
$P(S_{CON}|W)$: Consolidation model
$P(S|T)$: Translation model
$P(T)$: Language model (in target language)

Currently, speech translation for unrestricted domain has been much more intensively researched in the Global Autonomous Language Exploitation (GALE) projects funded by DARPA, which attempts to combine

speech recognition technologies with large speech and transcription corpora developed in the DARPA Speech Recognition Workshop and text translation technologies and parallel corpora developed in the TIDES projects. The GALE project has made publicly available huge corpora. However, 28% *MPAccy* for the MT of the manual transcription in our results indicates that translating Chinese speech into English text is still a very difficult, even if we can use large bilingual news corpora and the large speech data and manual transcriptions. To enhance the performance of speech translation, we need to solve problems in machine translation itself i.e., word reordering in translating language pairs with different syntactic structures and evaluation problems due to coverage of gold standards.

References

[1] K. Zechner, "Summarization of spoken language challenges, methods, and prospects," Speech Technology Expert eZine, no.6, 2002.
[2] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "A statistical approach for automatic speech summarization," EURASIP Journal on Applied Signal Processing, vol.2003, Issue 2, pp.128–139, 2003.
[3] C-STAR project, http://www.c-star.org/
[4] Verbmobile, http://verbmobil.dfki.de/
[5] NESPOLE! Poject, http://nespole.itc.it/
[6] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," IEEE Trans. Audio, Speech and Language Processing, vol.15, Issue 4, pp.1352–1365, 2007.
[7] DUC, http://www-nlpir.nist.gov/projects/duc/
[8] TIDES project, http://www.ldc.upenn.edu/TIDES/
[9] TREC, http://trec.nist.gov/data/qamain.html
[10] M. Cettolo, F. Brugnara, and M. Federico, "Advances in the automatic transcription of lectures," ICASSP, 2004.
[11] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," ASR, 2000.
[12] EARS project, http://www.nist.gov/speech/tests/rt/
[13] STR-DUST, http://www.is.cs.cmu.edu/str-dust/
[14] IWSpS, http://www.is.cs.cmu.edu/iwsps2004/
[15] M. Honal and T. Schultz, "Automatic disfluency removal on recognized spontaneous speech –Rapid adaptation to speaker dependent disfluencies," ICASSP, 2005.
[16] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," Eurospeech, 1997.
[17] TED corpus, http://www.elda.org/catalogue/en/speech/S0031.html
[18] H. Yu, Y.-C. Tam, T. Schaaf, S. Stüker, Q. Jin, M. Noamany, and T. Schultz, "The ISL RT04 mandarin broadcast news evaluation system," EARS Rich Transcription Workshop, 2004.
[19] C. Hori and A. Waibel., "Spontaneous speech consolidation for spoken language applications," Proc. Interspeech 2005, 2005.
[20] S. Vogel, "PESA: Phrase pair extraction as sentence splitting," Proc. MT Summit X, 2005.
[21] B. Zhao and S. Vogel, "A generalized alignment-free phrase extraction," Proc. ACL 2005 Workshop on Building and using Parallel Texts: Data Driven Machine Translation and Beyond (ACL WPT-05), June 2005.
[22] C. Hori, B. Zhao, S. Vogel, and A. Waibel., "Consolidation based speech translation," Proc. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU) 2007, 2007.
[23] GALE project, http://www.darpa.mil/ipto/programs/gale/gale.asp
[24] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," Comput. Speech Lang., vol.14, no.4, pp.373–400, 2000.
[25] N. Bertoldi, R. Zens, and M. Federico, "Speech translation by confusion network decoding," Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol.4, pp.IV-1297–IV-1300, April 2007.

Appendix A: Algorithm of Confusion Network Based Consolidation

First we define a notation used in the algorithm:

f, g, h :	A partial consolidated sentence hypothesis that has members of the score (*score*), the word sequence (*words*), and the position of the confusion set that the last word of the hypothesis is included (*pos*).
F,G,H,H′:	Hypothesis list that contains hypotheses.
$\hat{H}$:	Hypothesis list that contains complete hypotheses.
$\hat{h}$:	The best consolidated sentence hypothesis.
Generate():	function that generates a new hypothesis.
Insert(H, h):	Function that inserts h into H.
Move(H, F):	Function that moves all hypotheses in F to H.
ExpandHypo(h):	Function that generates a list of new hypotheses by adding each word that can succeed h.
CFNet(n):	Function that returns the n-th confusion set of words in the confusion network.

Second we describe the main procedure of the algorithm:

```
// Main procedure
begin
  h := Generate( )
  h. words := "<s>"
  h.pos := 1
  h.score := 0
  Insert(H, h)
  while H is not empty do begin
    foreach h ∈ H do begin
      F := ExpandHypo(h)
      foreach f ∈ F do begin
        if f.pos = N then // Is f a complete hypo?
          Insert(Ĥ, f)
        else
          Insert(H′, f)
      end
```

```
      end
      H:= H'
      H' := φ        // clear all hypotheses in H'
    end
    ĥ := max_{h∈Ĥ} h.score
end
```

$\hat{h}.words$ is the most likely consolidation result. For simplification, a pruning step is omitted in the above description.

Finally we show the procedure of *ExpandHypo*(h) that generates a list of new hypotheses according to the current hypothesis h and a given confusion network:

```
function ExpandHypo( h )
begin
  for n:=h.pos+1 to N do begin
    foreach w ∈ CFNet(n) do begin
      f := Generate( )
      f.pos := n
      if w = "@" then
        f.words := h.words
        f.score := h.score + λ_C C(n,"@")
        F:= ExpandHypo(f)
        Move(G, F)
      else
        f.words := h.words + w
            f.score := h.score + λ_L L(w|h.words)
                       + λ_C C(n, w) + sp*d(h, w) + ip
        Insert (G, f)
      endif
    end
  end
  return G
end
```

where the confidence score $C(w)$ is extended to $C(n,w)$ for using a confusion network, that indicates a logarithmic value of a posterior probability for word w in the n-th confusion set; $d(h,w)$ is a function that returns the number of skipped words between the last word of h and word w.

To improve search efficiency, in *Insert*(H,h) and *Move*(H,F), redundant hypotheses can be removed from the list. If there are multiple hypotheses which have reached the same position and whose last two words are identical, it is enough to retain only one hypothesis which has the maximum score among them in the list. For finding only the best complete hypothesis, it is not necessary to keep such redundant hypotheses. Since a trigram probability applied to the next word of the current hypothesis depends only on the last two words of the hypothesis, only the best hypothesis in the two-word context has a chance to be the best complete hypothesis i.e. the consolidation result in the future.

Appendix B: Example of Gradual Summarization

第十五届中美商贸联委会举行双方签署八项协议和换文。
The Fifteenth China-US Commerce and Trade Coordinative Commission was held with eight agreements signed and notes exchanged by the two parties.

中美商贸联委会举行双方签署八项协议和换文。
The China-US Commerce and Trade Coordinative Commission was held with eight agreements signed and notes exchanged by the two parties.

中美商贸联委会举行双方签署协议和换文。
The China-US Commerce and Trade Coordinative Commission was held with agreements signed and notes exchanged by the two parties.

中美商贸联委会举行签署协议和换文。
The China-US Commerce and Trade Coordinative Commission was held with agreements signed and notes exchanged.

中美商贸联委会举行签署协议。
The China-US Commerce and Trade Coordinative Commission was held with agreements signed.

中美商贸联委会举行。
The China-US Commerce and Trade Coordinative Commission was held.

中美商贸联委会。
The China-US Commerce and Trade Coordinative Commission.

Chiori Hori received the B.E. and the M.E. degrees in electrical and information engineering from Yamagata University, Yonezawa, Japan in 1994 and 1997, respectively. From April 1997 to March 1999, she was a Research Associate in the Faculty of Literature and Social Sciences, Yamagata University. In April 1999, she started the doctoral course in the Graduate School of Information Science and Engineering at Tokyo Institute of Technology (TITECH) and received her Ph.D. degree in March 2002. She was a Researcher in NTT Communication Science Laboratories (CS Labs) at Nippon Telegraph and Telephone Corporation (NTT), Kyoto, Japan from April 2002 to March 2004. She was a visiting researcher at Carnegie Mellon University in Pittsburgh from April 2004 to March 2006. She is currently a senior researcher at ATR and NiCT. She has received the Awaya Kiyoshi Science Promotion Award from the Acoustical Society of Japan (ASJ) in 2001 and Paper Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2003. She is a member of the IEEE and the ASJ.

Bing Zhao is a research staff member at IBM T.J. Watson Research. He got his Ph.D. in Language Technologies Institute, School of Computer Science, in Carnegie Mellon University in 2007. Before that, He got his master degree in artificial intelligence in National Lab of Pattern Recognition, Institute of Automation, in Chinese Academy of Sciences, and a bachelor degree in University of Science and Technology of China. His general research interests include the development of statistical machine translation and machine learning techniques for effective and efficient modeling of bilingual translational equivalences. He focuses on text translation, speech translations and general machine learning algorithms for natural language processing.

Stephan Vogel received the B.E. of physics at the Philipps University Marburg and the MPhil in History and Philosophy of Science from the University of Cambridge. He was a research assistant at the Lehrstuhl IV, Technical University (RWTH) of Aachen from 1995 to 2000. He has started to research at InterACT Lab in Karlsruhe University and Carnegie Mellon University since 2000. He is a research scientist in the Language Technologies Institute, School of Computer Science, Carnegie Mellon University. His research interest focus on machine translation.

Alex Waibel is a Professor of Computer Science at Carnegie Mellon University, Pittsburgh and at the University of Karlsruhe, Germany. He directs InterACT, the International Center for Advanced Communication Technologies at both Universities with research emphasis in speech recognition, language processing, speech translation, multimodal and perceptual user interfaces. At Carnegie Mellon, he also serves as Associate Director of the Language Technologies Institute and holds joint appointments in the Human Computer Interaction Institute and the Computer Science Department. He was one of the founders of C-STAR, the international consortium for speech translation research and served as its chairman from 1998-2000. His team has developed the JANUS speech translation system, the first American and European Speech Translation system, and more recently the first real-time simultaneous speech translation system for lectures. His lab has also developed a number of multimodal systems including perceptual Meeting Rooms, Meeting recognizers, Meeting Browser and multimodal dialog systems for humanoid robots. He directed the CHIL program (FP-6 Integrated Project on multimodality) in Europe and the NSF-ITR project STR-DUST (the first domain independent speech translation project) in the US. In the areas of speech, speech translation, and multimodal interfaces he holds several patents and has founded and co-founded several successful commercial ventures. He received the B.S. in Electrical Engineering from the Massachusetts Institute of Technology in 1979, and his M.S. and Ph.D. degrees in Computer Science from Carnegie Mellon University in 1980 and 1986. His work on the Time Delay Neural Networks was awarded the IEEE best paper award in 1990. His contributions to multilingual and speech translation systems was awarded the "Alcatel SEL Research Prize for Technical Communication" in 1994, the "Allen Newell Award for Research Excellence" from CMU in 2002, and the Speech Communication Best Paper Award in 2002.

Hideki Kashioka received his Ph.D. in Information Science from Osaka University in 1993. From 1993, he works for ATR Spoken Language Translation Research Laboratories. He is currently the head of department of Spoken Language Research in ATR Spoken Language Communication Research Laboratories. He is also the research manager of Spoken Language Communication Group at Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology. he is also the visiting associate professor of the graduate school of Information Science at the Nara Institute of Science and Technology from 1999. He is a member of ANLP, JCSS, JSAI and IPSJ.

Satoshi Nakamura was born in Japan on August 4, 1958. He received his B.S. in Electronic Engineering from the Kyoto Institute of Technology in 1981 and his Ph.D. in Information Science from Kyoto University in 1992. From 1981 to 1993, he worked for Sharp's Central Research Laboratory in Nara. From 1986 to 1989, he worked for ATR Interpreting Telephony Research Laboratories. From 1994 to 2000, he was the associate professor of the graduate school of Information Science at the Nara Institute of Science and Technology. In 1996, he was a visiting research professor of the CAIP center at Rutgers University in New Jersey. He is currently the director of ATR Spoken Language Communication Research Laboratories. He is also the director of the MASTAR project at Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology. He also serves as an honorary professor of University Karlsruhe, Germany since 2004. He received the Awaya Award from the Acoustical Society of Japan in 1992, the Interaction 2001 Best Paper Award in 2001, Yamashita Research Award from the Information Processing Society of Japan in 2001, Telecom System Award, AAMT Nagao Award and Docomo Mobile Science Award in 2007. He served as an associate editor for the Journal of the IEICE ED from 2000 to 2002, a member of the Speech Technical Committee of the IEEE Signal Processing Society in 2001-2004, a general chair of International Workshop of Spoken Language Translation (IWSLT2006) and Oriental Cocosda 2008, and a technical chair IEEE ASRU2007 and INTERSPEEC 2010. He is a member of IEEE, IPSJ and ASJ.