INVITED PAPER *Special Section on Award-winning Papers*

Development of the "VoiceTra" Multi-Lingual Speech Translation System

Shigeki MATSUDA[†*a)], ***Member***, **Teruaki HAYASHI**[†*], **Yutaka ASHIKARI**[†], **Yoshinori SHIGA**[†], ***Nonmembers***, **Hidenori KASHIOKA**[†], **Keiji YASUDA**[†**], ***Members***, **Hideo OKUMA**[†***], **Masao UCHIYAMA**[†], ***Nonmembers***, **Eiichiro SUMITA**[†], **Hisashi KAWAI**[†], ***and*** **Satoshi NAKAMURA**[†****], ***Members***

SUMMARY This study introduces large-scale field experiments of VoiceTra, which is the world's first speech-to-speech multilingual translation application for smart phones. In the study, approximately 10 million input utterances were collected since the experiments commenced. The usage of collected data was analyzed and discussed. The study has several important contributions. First, it explains system configuration, communication protocol between clients and servers, and details of multilingual automatic speech recognition, multilingual machine translation, and multilingual speech synthesis subsystems. Second, it demonstrates the effects of mid-term system updates using collected data to improve an acoustic model, a language model, and a dictionary. Third, it analyzes system usage.

key words: *speech translation, statistical machine translation, speech recognition, speech synthesis*

1. Introduction

Speech translation technologies realize smooth communication between people from different parts of the world who speak different languages. The demand for these technologies has steadily increased in the last few years with the spread of the Internet and globalization of the economy [1], [2]. This has triggered research and development activities related to speech translation systems in several countries around the world.

In Japan, both the Advanced Telecommunications Research Institute International (ATR) and the National Institute of Information and Communications Technology (NICT) are engaged in the research and development activities of speech translation systems [4] through field experiments by assuming actual use environments [2], [3]. Additionally, the Verbmobil project [5] in Germany conducts research on speech translation between Japanese, English, and German, and the NESPOLE! project in the EU [6] and TC-STAR [7] conduct research on speech translation technologies utilized in international conferences to translate lectures. In the United States, active research on speech translation technologies was conducted by the TransTac project as well as the GALE project [8] of the Defense Advanced Research Projects Agency (DARPA). These projects focused specifically on translating languages such as Arabic and Chinese to English with the aim of extracting information. In Asia, eight countries (Japan, China, Korea, Thailand, Indonesia, Malaysia, Vietnam, and Singapore) that are collectively termed the Asian Speech Translation Advanced Research Consortium (A-STAR) [9] are engaged in collaborative research and development activities on speech recognition, translation, and text-to-speech technologies, and collection of speech and text corpora. Recently, an international research consortium called U-STAR [10], which includes 26 institutes from 23 different countries (as of Oct., 2013) including NICT, has collaboratively developed a network-based speech-to-speech translation system by mutually connecting speech recognition, machine translation, and text-to-speech servers developed by each member of the consortium via the network.

This study describes "VoiceTra," which is the world's first network-based speech-to-speech translation system that is operated on smart phones, and a large-scale field experiment conducted using VoiceTra. The experiment was launched in July 2010 with the aim of analyzing collected speech data and improving the system using the collected speech data. The VoiceTra system consists of a client app that is capable of displaying the speech recognition/translation results and playing synthesized speech. It also includes servers that handle speech recognition, translation, and text-to-speech processes. The client app was first developed for Apple's iPhone and is available on the App Store for free. The Android version was released in April 2011 to reach a wider audience and is also available for free on the Android Market. The apps together comprised 700,000 downloads and over 10 million accesses by the end of 2012. This study details the system configuration of VoiceTra as well as the multilingual speech recognition, translation, and text-to-speech technologies utilized by the system. To review user statistics, the analysis results of the collected large-scale real use speech data that were created by actually monitoring and analyzing the data by hand are discussed. The collected data were used to apply unsupervised adaptation on the acoustic and language models for speech recognition as part of the experiment. The study also discusses the improvements involved in speech recognition and translation accuracies with the application of unsupervised adaptation.

Manuscript received November 7, 2016.
Manuscript revised December 22, 2016.
Manuscript publicized January 13, 2017.
[†]The authors are with National Institute of Information and Communications Technology, Kyoto-fu, 619–0289 Japan.
[*]Presently, with ATR-Trek Co., Ltd.
[**]Presently, with KDDI Research, Inc.
[***]Presently, with FEAT Ltd.
[****]Presently, with Nara Institute of Science and Technology.
a) E-mail: shigeki.matsuda@atr-trek.co.jp
DOI: 10.1587/transinf.2016AWI0006

This paper is organized as follows. Section 2 gives an overview of VoiceTra and its user interface, communication protocol between the client and the server, and system configuration of actual servers. Section 3 describes multilingual speech recognition, translation, and text-to-speech technologies used in VoiceTra. Section 4 discusses manually monitored and analyzed results of the collected speech data and evaluates speech translation accuracies. Finally, Sect. 5 presents the conclusions.

2. "VoiceTra" Network-Based Multi-Lingual Speech-to-Speech Translation System

2.1 System Overview

VoiceTra is a network-based multi-lingual speech-to-speech translation system realized using servers for speech recognition, language translation, and speech synthesis prepared on a network such as the Internet. Translation is performed by processing speech on each server in that order, and the speech is translated into a foreign language. These processes typically require large amounts of calculation on the server and therefore could utilize large-scale models for speech recognition and language translation. Therefore, VoiceTra realizes speech translation with higher performance than a stand-alone system that performs all processes on a smart phone.

Figure 1 shows the structure of the VoiceTra system. The VoiceTra client works for speech input, displays a translation result, and plays the synthesized speech. The VoiceTra server consists of multiple servers wherein each server performs a specific process such as speech recognition, language translation, and speech synthesis. Speech recognition and speech synthesis servers are prepared for each language, and a language translation server is prepared for each language pair. Hence, it is possible to update a model and server software for a specific language or language pair without shutting down all servers. Table 1 shows a list of languages that can be used in VoiceTra. As shown in this table, the VoiceTra client can access speech recognition and speech synthesis servers for six languages: Japanese, English, Chinese, Indonesian, Vietnamese, and Korean. Language translation servers for 210 language pairs between 21 languages are available. Additionally, text input and output is available for a language not supporting speech input and output. Text input and output continues to be an important function to communicate with foreigners in situations where speech recognition and speech synthesis are not available such as very noisy environments or libraries.

2.2 Communication Protocol between the Client and the Server

Speech Translation Markup Language (STML) [11] is used to communicate between the client and the server. Following a speech input from the client, three processes, namely

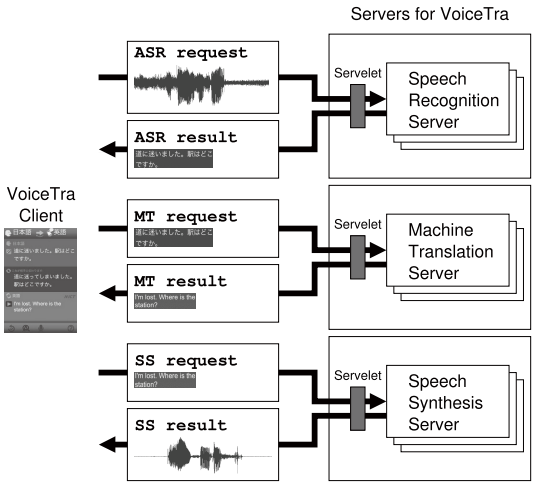

Fig. 1 Structure of the VoiceTra system and data communication between client and server

Table 1 List of available languages that can be translated by VoiceTra

Available to use speech input and output	Available to use text input and output
Japanese	Japanese
English	English
Chinese	Chinese
Indonesian	Indonesian
Vietnamese	Vietnamese
Korean	Korean
	Taiwanese Mandarin
	French
	German
	Hindi
	Italian
	Malay
	Portuguese
	Portuguese (Brazil)
	Russian
	Spanish
	Tagalog
	Thai
	Arabic
	Dutch
	Danish

speech recognition, language translation, and speech synthesis, are executed by receiving a request of speech recognition from the client <SR_IN>, sending a recognition result from the server <SR_OUT>, requesting language translation from the client <MT_IN>, sending a translation result from the server <MT_OUT>, requesting speech synthesis from the client <SS_IN>, and sending synthesized speech from the server <SS_OUT> via the network.

Multipurpose Internet Mail Extensions (MIME) are used to convert necessary information for speech translation (such as source language, target language, and speech waveform) to a text, and then a Hypertext Transfer Protocol (HTTP) is used to transfer the text.

A speech translation system can be easily realized by using a single server in which speech recognition, language translation, and speech synthesis are performed. However,

it is easy to extend STML to a language or language pair for speech recognition, language translation, and speech synthesis. Specifically, STML was proposed as a protocol to realize multilingual speech translation by integrating multiple servers in several companies and research institutions for speech recognition, language translation, and speech synthesis. Hence, it is possible to perform speech translation using these servers by only registering information including newly added speech recognition, language translation, and speech synthesis servers to the client. Hence, STML was adopted in this study from the viewpoint of future expansions of a potential translation language.

1. Request of speech recognition <SR_IN>
   Information necessary for speech recognition, such as maximum number of N-best (MaxNBest), language of input speech (Language), speech codec (Audio), and sampling frequency (SamplingFrequency), is described in STML format and transferred to a speech recognition server. Utterances spoken to the VoiceTra client are recoded by 16 kHz 16-bit sampling and compressed by ADPCM to 1/4. This is followed by transferring the compressed speech data to the server. In a normal use scenario, the maximum number of N-best is set as one. Therefore, language translation and speech synthesis are performed for a recognition result of 1-best.
2. Sending the recognition result <SR_OUT>
   The STML format is used to describe the word sequence (NBest) obtained by speech recognition, and then the recognition result is transferred to the client. Additionally, UTF-8 is used for the character coding.
3. Requesting language translation <MT_IN>
   The STML format is used to describe the source language of translation (SourceLanguage), target language (TargetLanguage), and word sequence (NBest) in the source language. Following this, the language translation request is transferred to a language translation server.
4. Sending the translation result <MT_OUT>
   The STML format is used to describe the source language of translation (SourceLanguage), target language (TargetLanguage), and word sequence (NBest) in the target language, and then the translation result is transferred to the client.
5. Requesting speech synthesis <SS_IN>
   The STML format is used to describe the speech codec (Audio), sampling frequency (SamplingFrequency), language for speech synthesis (Language), and word sequence (NBest). This is followed by transferring the speech synthesis request to the speech synthesis server.
6. Sending synthesized speech <SS_OUT>
   The STML format is used to describe the speech codec (Audio), sampling frequency (SamplingFrequency), and language for speech synthesis (Language), and then the sent synthesized speech is transferred to the client. In a manner similar to <SR_IN>, the synthesized speech waveform is transferred by ADPCM to the client.

Fig. 2 Translation panel (left) and language selection panel (right)

2.3 User Interface

Figure 2 shows a screen-shot of VoiceTra. The left side corresponds to a translation screen, and the right side corresponds to the language selection screen. The screen shown in this example involves a translation from Japanese to English. The speech input is timed to commence when the user brings the iPhone close to his/her ear and to stop when the iPhone is placed down. The vibration of the terminal indicates the commencement of speech input to the user. The application of proximity sensors to disable touch panel operation while talking made such an interface possible. Although this speech input interface corresponds to a natural action for the user, it is possible to forcibly close the distance between the microphone and the mouth. Thus, a relatively high signal-to-noise ratio (SNR) is obtained.

The top row displays the speech recognition results of a Japanese utterance that corresponds to "道に迷いました。駅はどこですか。," and the bottom row displays the translation results that correspond to "I'm lost. Where is the station?" The synthesized speech is played from the iPhone following the display of the translation results. The Japanese characters in the middle row indicate the English translation result that was translated back into Japanese, i.e., back translation [12]. The user can verify the accuracy of the translation against the speech recognition result by comparing these back-translation results with the meaning of the speech recognition results.

2.4 Speech Translation Server

It is necessary for a speech translation system to translate a user's speech in real time. Additionally, it is desirable to obtain the speech translation result in a very short time with respect to the end of the utterance. Decoding parameters, such as beam width, were adjusted to the real time factor (RTF) that is 1.0 or less to ensure that the speech recognition result was obtained at the time when the utterance ends. The RTF of the language translation process corresponded to 0.05. Finally, the speech translation system was adjusted

Table 2 Number of servers for language translation

	Target lang.				
Source lang.	Japanese	English	Chinese	Korean	Others
Japanese	–	9	9	9	3
English	9	–	3	3	3
Chinese	9	3	–	3	3
Korean	9	3	3	–	3
Others	1	1	1	1	1

Table 3 Corpus size for building acoustic models

Language	Utterances	Hours	Speaking style
Japanese	227k+60k	390.0 + 57.5	read speech+field speech
English	256k	267.3	read speech
Chinese	495k	509.8	read speech
Indonesian	84k	79.3	read speech
Vietnamese	23k	19.4	radio speech
Korean	149k	271.3	read speech

so that the translation result was obtained within 1.05 times the utterance length with the exception of a network delay.

The specs of the required server were estimated for the general public VoiceTra. It was assumed that each of 0.5M clients utilized approximately 10 s per week. Furthermore, it was assumed that two times the access of the above assumption occurred at the time of the peak. Given these assumptions, a speech of 10M s (= 0.5M clients × 10 s × 2) would be translated in a week. The RTF was adjusted to approximately 1.0, and thus a CPU time of 10M s was also required to translate a speech of 10M s. Given that a week has approximately 0.6M s, 16.7 CPU cores (= 10M/0.6M) were required to satisfy these assumptions. This is equivalent to a process involving a speech of 1000 s per min.

Six servers, in which each server included four CPU cores, were used in the actual operation, and thus a total number of 24 cores existed. The numbers of speech recognition, language translation, and speech synthesis servers for each language were determined from the ratio of the actual access. The number of speech recognition servers for Japanese and English was nine and four, respectively, and the number of speech recognition servers for Chinese, Indonesian, Vietnamese, and Korean was two. The number of speech synthesis servers was equivalent to the number of speech recognition servers for each language. Table 2 shows the number of language translation servers. As shown in the table, there were more than two language translation servers for language pairs involving Japanese, English, Chinese, and Korean to continue the service, even in the event of failure of a server.

The actual average access time was 25.8 s per min. However, when VoiceTra was introduced in a TV program, a delay occurred in the speech translation process because it was necessary to process speech data up to 1551.1 s per min.

3. Multilingual Speech-to-Speech Translation System

3.1 Multilingual Speech Recognition

The speech recognition system used in VoiceTra consisted of a frontend, which performed noise suppression using particle filtering [13] and acoustic analysis, and a backend, which performed large vocabulary continuous speech recognition (LVCSR) using ATRASR [14].

A feature vector consisted of 12 MFCCs, 12 Δ MFCCs, and Δ pow extracted from a 20-ms window size with a 10-ms frame shift for a speech waveform recorded by a sampling frequency of 16 kHz. Two-pass Cepstrum Mean Subtraction (CMS) is widely used to normalize the channel characteristics of a microphone. However, while using a two-pass CMS, it is necessary to wait for the utterance to end in order to start the recognition process. This is unsuitable for a VoiceTra system in real time. Additionally, it is not possible to use one-pass CMS using the average cepstrum of the previous utterance because it is difficult to store the last speech of each of the terminals on the server side. In the system proposed in the study, a successive channel characteristics normalization using a prior distribution given by the following equation was used.

$$s_t = \frac{\tau s^{pri} + \sum_{u=1}^{t} c_u^{org}}{\tau + t} \quad (1)$$

$$c_t^{cms} = c_t^{org} - s_t \quad (2)$$

Here, c_t^{org} and c_t^{cms} denote cepstrum vectors before and after normalization at time t, respectively. Furthermore, s^{pri} denotes a prior distribution, and the average vector calculated from a large amount of speech data collected by terminals supported by VoiceTra, such as an iPhone, is used as the prior distribution. Additionally, τ denotes the weight for the prior distribution, and s_t denotes the average cepstrum used for normalization at time t.

The Japanese acoustic model involved approximately 390 h of read speech in which approximately 4,500 adult speakers were used. Gender-dependent triphone hidden Markov models (HMMs) with 5,660 states were estimated using the read speech that was contaminated by noise sources, such as train stations, department stores, restaurants, and cars, at 10–30 dB SNR. Each state had 10 mixture components. Finally, the Japanese acoustic model was estimated by the MAP adaptation [15] using the speech collected in the field experiments conducted in five regions in Japan [2]. The acoustic models for English, Chinese, Korean, and Indonesian were estimated using read speech. The acoustic model for Vietnamese was estimated using radio speech and related transcriptions. In a manner similar to Japanese, gender-dependent triphone HMMs for these languages were estimated using a speech corpus that was contaminated by several types of noise at 10–30 dB SNR. Table 3 shows the corpus size used for building acoustic models.

Multiclass composite bi-gram [16] and word tri-gram language models were estimated using a text corpus collected from travel conversations. Table 4 shows the corpus size required to estimate the language model for each language. Japanese, English, and Chinese language models

Table 4 Corpus size for building language models

Language	# of utterances	# of words	Lexicon size
Japanese	803k	8,154k	63k
English	703k	5,967k	44k
Chinese	715k	4,266k	45k
Indonesian	170k	1,121k	15k
Vietnamese	162k	1,432k	8k
Korean	330k	2,948k	43k

Table 5 Corpus size for building machine translation systems

Language pair	# of sentences	# of words	Lexicon size
Japanese, English	701k	6,350k	43k
Japanese, Chinese	508k	4,035k	33k
Japanese, Korean	392k	2,964k	29k
Japanese, Other languages	160k	1,180k	16k

were estimated with sentences from the Basic Travel Expression Corpus (BTEC) [20] and the transcribed text of the speech collected in the field experiments involving five regions in Japan. Indonesian, Vietnamese, and Korean language models were estimated using only the BTEC. The lexicon size of the Vietnamese language model was smaller than those of other languages because the lexicon consists of phonemes and syllables [17].

3.2 Multilingual Machine Translation System

The machine translation part consisted of a translation memory and a statistical machine translation system. The translation memory yielded output when the input sentence exactly matched a sentence in the parallel corpus. A phrase based statistical machine translation system was introduced, and it worked only when the translation memory had no output. The system involved eight features [18] including source language to target language phrase unit translation probability (translation model) and target language side 5-gram language model.

Each model was trained using a Moses toolkit [19]. Table 5 shows the details of the training corpus. The corpus was used for both translation memory and statistical machine translation training. As shown in the table, the training set size was the largest in the Japanese and English language pair. The size corresponded to 700,000 sentence pairs, including the ATR travel conversation corpus [21] and the BTEC. Subsets of the Japanese-English training set were manually translated to obtain the training sets of the other language pairs.

3.3 Multilingual Text-to-Speech Synthesis

The text-to-speech synthesis system was composed of two modules: a text processing module and a speech signal processing module. The former converted input text into a sequence of context-dependent phone labels by looking up pronunciation lexicons and applying a set of rules developed to analyze text in each language. The context-dependent phone label included various types of linguistic information such as the phonetic type and phonetic context of the current phone, the part-of-speech of the word that the phone belonged to, the dependency of the word, and the positions of the phone within the sentence, clause, and phrase. All the fore-mentioned types of linguistic information were used in the subsequent speech signal processing module to control the phonetic and prosodic aspects of speech to be synthesized. With respect to languages involving a written text with no spaces between words, such as Chinese and Japanese, the input text was segmented into morphemes prior to the fore-mentioned text-to-label-sequence conversion to determine boundaries of words/phrases in the text. The morphological segmentation of such text in Japanese and Chinese involved the use of open-source analyzers "Chasen" [23] and "Mecab" [24], respectively.

Table 6 Size of the corpus that was used to build the speech synthesis model of each language

Language	Number of utterances	Total duration of speech in hours
Japanese	19k	25.0
English	18k	17.4
Chinese	15k	20.3
Indonesian	2k	1.9
Vietnamese	1k	0.6
Korean	4k	8.9

The speech signal processing module then synthesized speech according to the context-dependent labels. This process involved the use of a statistical method based on HMMs [25], [26]. Among several types of HMMs, hidden semi-Markov models (HSMMs) with a five-state left-to-right topology with no skip were utilized in the present study. The models were trained in advance with acoustic features that were extracted every 5 ms from prerecorded speech sampled at 16 kHz. The features were composed of a 39-dimensional mel-cepstral vector, logarithmic fundamental frequency (F_0), and dynamic features (Δ and Δ^2). The mel-cepstral vectors were computed from the spectral envelopes obtained by the STRAIGHT analysis [27]. Table 6 lists the size of the speech corpus used to train the models for each of the languages.

4. Field Experiments

4.1 Analysis of Collected Real-Use Speech Data

Figure 3 shows the cumulative number of accesses to the VoiceTra server between July 2010 and January 2013. As is evident from the figure, the number of accesses steadily increased since the launch of the service. Approximately 10.54 million utterances were collected by the end of 2012. The breakdown of the utterances by language is as follows: Japanese, English, Chinese, Korean, Indonesian, and Vietnamese utterances accounted for 74%, 19%, 4%, 2%, 1%, and 1% of the total utterances, respectively. Accesses from within and outside Japan corresponded to 94.7% and 5.3%, respectively, and thus most of the accesses were from within Japan. Figure 4 shows the country-wise breakdown of the

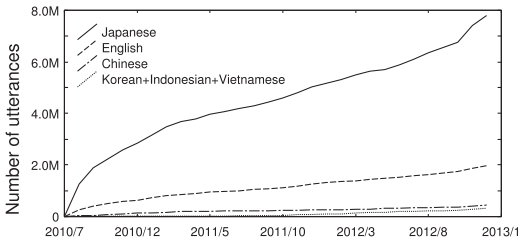

Fig. 3 Total number of accesses to the VoiceTra server

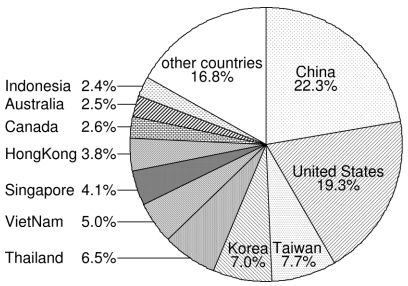

Fig. 4 Proportion of accesses for each country

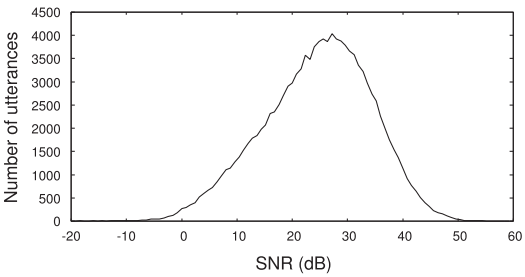

Fig. 5 Histogram of SNR on VoiceTra speech data

overseas accesses. The accesses were mostly from China (22.3%), followed by the United States (19.3%), Taiwan (7.7%), and Korea (7.0%). Accesses from Indonesia and Vietnam contributed to 2.4% and 5.0% of the utterances, respectively, given that VoiceTra supported voice input in Indonesian and Vietnamese.

Figure 5 shows the distribution of SNR that was computed from speech collected in the field experiments. The SNR was obtained by first determining speech-with-noise and non-speech sections according to the results of phone alignment performed using annotated text and then calculating the ratio of the mean power of speech between these sections. As observed in the figure, the SNR exhibited a wide range from 0 dB to 50 dB with a median of 25 dB at the center. This revealed that it was required to deal with speech of a relatively low SNR.

To review the usage statistics, utterances collected from 478 terminal devices that were most frequently used within Japan during the experiment were classified into several types of user purposes predicted by monitoring the utterances. Each of the devices was among the top 100 devices for any of the following assessments: total frequency of use, frequency of use per day, number of months, number

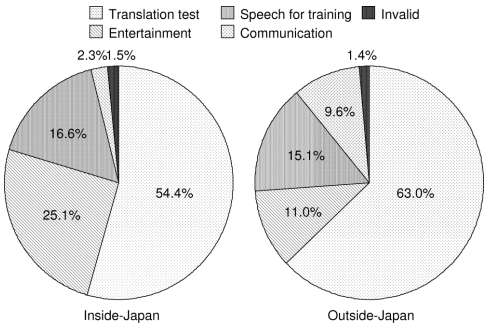

Fig. 6 Types of utterances collected using VoiceTra on inside- and outside-Japan

of days, and number of consecutive days they were used for. The left-hand part of Fig. 6 shows the classification. In the chart, "Translation test" indicates utterances that were judged as given when the user was checking translation results beforehand by him/herself or when the user was simply checking a word in the dictionary. "Entertainment" indicates utterances that were spoken just for fun, such as utterances from famous cartoon animations and song lyrics. "Speech for training" indicates utterances in English that were by Japanese individuals for the purpose of checking speech recognition results. "Communication" indicates utterances involved in real conversations with a foreigner. "Invalid" implies that no utterance could be found in the input.

As shown in the figure, the most common utilization purpose corresponded to "Translation test" and accounted for 54.4%. This was followed by "Entertainment" in the second place (25.1%). This in turn was followed by "Speech for training" (16.6%) and "Communication" (2.3%). With respect to the use within Japan, practical use as a speech translator corresponded to a total of 56.7% including the sum of proportions of "Translation test" as well as "Communication." This was followed by classifying utterances collected from 146 terminal devices used outside Japan in accordance with the different types of purposes. Each of the devices was among the top 50 devices for any of the same assessments above. The right-hand part of Fig. 6 shows the classification. As shown in the chart, utterances collected in actual communication with foreigners accounted for 9.6% and indicated a higher percentage when compared with the usage within Japan. The percentage of cases wherein the application was practically used as a speech translator corresponded to 71.0%. Among the 14 terminal devices that were interpreted as used for verbal communication, a device was used by a native speaker of Chinese and the rest by native speakers of Japanese. According to the contents of the utterances, the Japanese speakers appeared to have used VoiceTra to converse with local people during their sightseeing/business trips. There was a greater need to communicate with people who did not speak Japanese overseas, and this was probably the reason for the increase in the proportion of use for communication purposes.

Table 7 Model combinations of three speech-to-speech translation systems

	2010/08	2010/11	2011/06
STS	STS_1	STS_2	STS_3
MT	MT_1	MT_1	MT_2
ASR	ASR_1	ASR_2	ASR_3
AM	AM_1	AM_2	AM_2
LM	LM_1	LM_1	LM_2

Table 8 The VoiceTra testset

Test set	Total	Travel related	Others
2010/08	700	587	113 (27)
2010/11	700	598	102 (29)
2011/06	700	599	101 (25)

4.2 Evaluation of Speech Translation Quality

This subsection explains the evaluation of speech translation quality using VoiceTra users' speech data. VoiceTra was released in end of July 2010. Following the first release, the system was updated using VoiceTra users' speech data as follows:

- Acoustic model adaptation on November 2010 (2010/11)
- Language model adaptation and lexicon expansion (adding 14,000 entries) for ASR and MT on June 2011 (2011/06)

Unsupervised adaptation for ASR including acoustic model adaptation and language model adaptation was performed to reduce transcription cost [29]–[31].

Table 7 shows the three system configurations of the Speech Translation System (STS). With respect to Machine Translation (MT), MT_1 indicates the baseline of the MT system explained in Sect. 3.2. Additionally, MT_2 indicates the lexicon expanded version. With respect to Automatic Speech Recognition (ASR), AM_1 and AM_2 indicate the baseline acoustic model explained in Sect. 3.1 and the unsupervised adaptation version acoustic model, respectively. Furthermore, LM_1 and LM_2 indicate the baseline language model and the unsupervised adaptation language model with 14,000 additional new words, respectively.

Three test sets were maintained by randomly selecting users' utterances from the VoiceTra log to evaluate the effect of unsupervised model adaptation and lexicon expansion. Each test set consisted of 700 utterances. The differences between the three test sets corresponded to the following data corrected periods: after the first release, after the first update, and after the second update. These test sets did not contain utterances with barnyard contents or utterances recorded by erroneous operations because these utterances were manually filtered in advance. To analyze test utterances, they were manually classified into two classes: travel-related utterances and not-travel related utterances. Table 8 shows the results. The numbers in brackets indicate the number of unclear utterances.

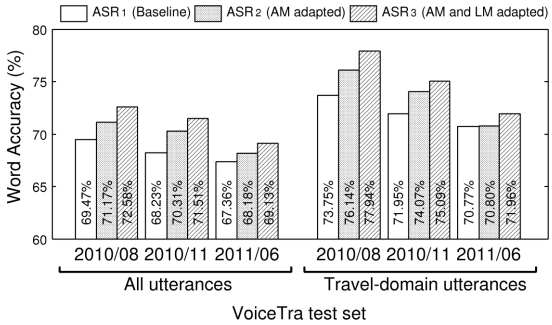

Fig. 7 ASR performance on the VoiceTra test set

Unsupervised acoustic model adaptation was performed using speech recognition results for speech waveforms collected by VoiceTra. MAP adaptation [31] was applied to mean vectors in individual Gaussian components using speech periods with higher word confidences [28] when compared with the threshold. Additionally, unsupervised language model adaptation was performed using only recognition results with sentence confidences that exceeded the threshold [31]. Specifically, 1,000,000 utterances collected on August 2010 (2010/08) were used for the unsupervised adaptation. Moreover, 14,000 words were added to the lexicon. With respect to the statistical machine translation system, a method proposed in a previous study was used to expand the bilingual dictionary size [22].

4.2.1 Evaluation of Speech Recognition Performance

The left half of Fig. 7 indicates the word accuracy rates of all utterances of the test set. As shown in the figure, the ASR system (ASR_2) with acoustic model adaptation achieved higher speech recognition performance than that of the baseline system (ASR_1). We have also confirmed that the system with both acoustic and language model adaptations (ASR_3) achieved even greater performance. The right half of Fig. 7 indicates the word accuracy rates of travel-domain utterances which were sorted by hand. As shown in the figure, the word accuracy rates of travel-domain utterances are higher than that of all utterances of the test set. Unsupervised acoustic model adaptation contributes to performance improvements, enabling adaptation to acoustic environments such as noise and adaptation to speaker variation and style. We recalculated the word accuracy rates of 767 utterances from the test set with SNR higher than 30dB, in order to see the ASR performance with relatively clean speech in speculation that they would have less effects on degradation due to noise. Word accuracy rates of the baseline ASR system (ASR_1) and the ASR system with acoustic model adaptation (ASR_2) were 77.19% and 78.69%, respectively. The ASR system with acoustic model adaptation (ASR_3) achieved higher word accuracy rates than that of the baseline system, even with the test set consisting of clean speech only. From these results, we have confirmed that improvements in ASR performance were achieved not only with adaptation to acoustic environments such as noise, but

Table 9 Perplexity of the VoiceTra test set with respect to the language models used for speech recognition

	LM_1		LM_2	
Test set	All	Travel	All	Travel
2010/08	74.6	45.2	50.9	30.9
2010/11	97.3	63.4	73.7	48.7
2011/06	98.0	68.4	74.1	52.0

Table 10 Number of out of vocabulary words in the VoiceTra test set with respect to the language models used for speech recognition

	LM_1		LM_2	
Test set	# of OOVs	OOV rate	# of OOVs	OOV rate
2010/08	55	1.7%	50	1.6%
2010/11	66	2.2%	61	2.0%
2011/06	69	2.3%	61	2.0%

with adaptation to speaker variation and style. Table 9 shows the perplexities for all utterances (All) and travel-domain utterances (Travel) of the VoiceTra test set in each of the calculated month. As shown in this table, the adapted language model (LM_2) achieved lower perplexities than that of the baseline language model (LM_1). The transcription text used for language model adaptation includes about 25% of greeting words such as "hello" and "good evening." It is conceivable that the improvements in perplexities using the adapted language model (LM_2) were obtained by the adaptation to word occurrence frequency.

Speech recognition performance degraded with each passing day. Table 9 shows the perplexity in the VoiceTra test-set, and Table 10 shows the number of Out of Vocabularies (OOVs) in conjunction with the OOV rates. As shown in the tables, these parameters (i.e., number of OOVs, OOV rates, and perplexity) increased with each passing day. When the VoiceTra service started, the proportion of simple sentences that consisted of few words, such as salutations, was high in terms of user utterances. However, speech recognition performance degraded given the increase in the complexity of user utterances with each passing day.

4.2.2 Evaluation of Translation Quality

This section describes the evaluation of translation quality. The same test sets as those in ASR evaluation were used. However, test utterances with unclear intended meanings were excluded from the translation quality evaluation. With respect to the subjective evaluation, the following five translation quality ranks were defined: S (Perfect), A (Correct), B (Fair), C (Acceptable), and D (Nonsense). Specifically, three versions of STS systems (STS_1, STS_2, and STS_3 in Table 7) and two versions of MT systems (MT_1 and MT_2 in Table 7) were evaluated. With respect to the STS evaluation, the input format corresponded to a speech signal. In contrast, the input format corresponded to manually transcribed text for the MT evaluation.

Figure 8 shows the evaluation results of the translation quality for all the test sets. Conversely, Fig. 9 shows the evaluation results for the travel-related test utterances. In

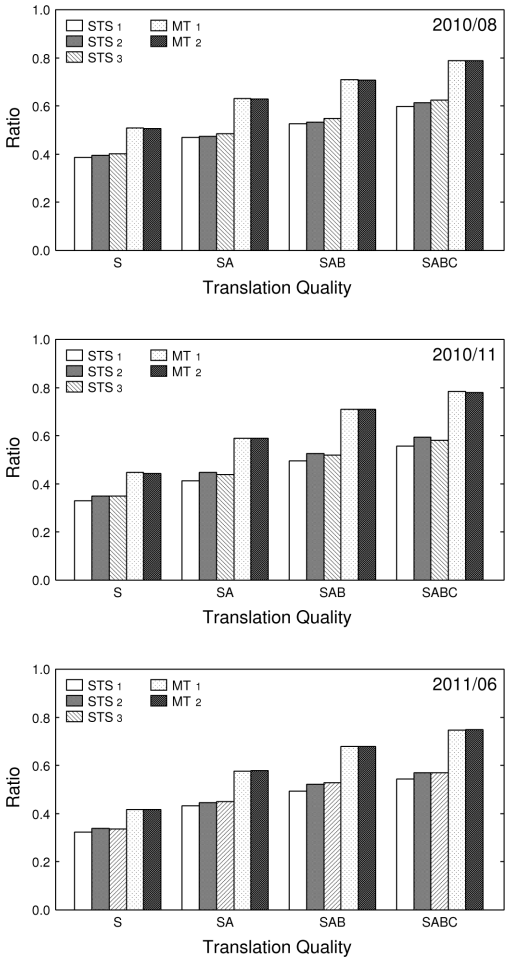

Fig. 8 Evaluation results of Japanese-to-English translation quality on the VoiceTra test set

these figures, the vertical axes indicate the ratio of test utterances to the test set. The labels on horizontal axes indicate the calculation method for the evaluation results. For example, SA indicates the ratio of test utterances in S and A translation quality rank to the test set.

Observations revealed that STS_2 and STS_3 exhibited better results when compared with those of STS_1. That is, ASR system updates improved STS system performance. A comparison of Figs. 8 and 9 indicated that the evaluation results using travel related utterances (62.3% given the condition of STS_3 and SABC in Fig. 9) produced better results when compared with the evaluation results using whole test utterances (59.2% given the condition of STS_3 and SABC in Fig. 8). This implied that 62.3% of STS output was at worst acceptable translation in travel related input.

This section discusses the effects of the MT update. As previously mentioned, only bilingual dictionary expansion was applied as the MT update. This type of update works well only if the added dictionary entries appear in the input sentence. The other case includes a possibility of producing a side effect that causes the degradation of translation quality. Table 11 shows the number of out of vocabulary

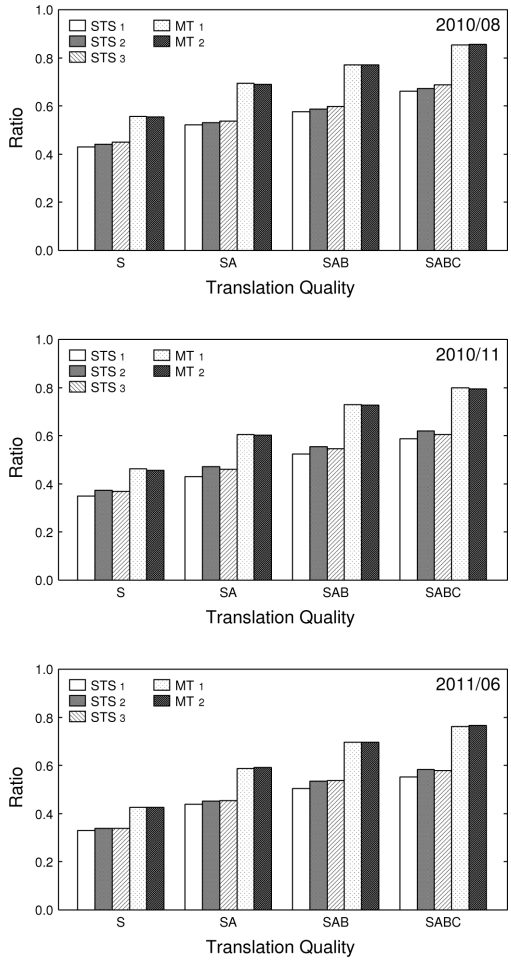

Fig. 9 Evaluation results of Japanese-to-English translation quality on travel-domain utterances in the VoiceTra test set

Table 11 Number of out of vocabulary words in the VoiceTra test set with respect to the dictionary in the translation sytem

Testset	MT_1	MT_2
2010/08	133	120
2010/11	146	142
2011/06	138	134

(OOV) words for each MT version, namely MT_1 and MT_2. As shown in the table, 21 words corresponded to non-OOV words by the MT update, and 21 test utterances containing these words were analyzed. The results indicated that 19 out of 21 utterances corresponded to D-rank translation quality or no output for MT_1. Meanwhile, MT_2 could yield over C-rank translation quality for 14 utterances from 21 utterances. Figure 9 indicates the degradation of MT_2 on the test set 2010/11. However, the degradation was very small. Thus, these evaluation results indicated that the MT update was effective and that the side effect was very small. Future research will include investigating methods to effectively add useful dictionary entries.

5. Conclusion

This study describes the "VoiceTra" system that was developed by NICT and released as the world's first network-based multilingual speech-to-speech translation system that was operated on a smart phone. The system configuration, communication protocol between clients and servers, and details of multilingual automatic speech recognition, multilingual machine translation, and multilingual speech synthesis subsystems are described. Analysis of system usage and speech translation performance evaluations are examined.

VoiceTra was downloaded 700,000 times and accessed approximately 10 million times as of December 2012. The usage analysis results confirmed that the practical use of VoiceTra as a speech translator corresponded to 56.7% within Japan and 72.6% outside Japan. Specifically, the percentage of actual communication with foreigners accounted for 9.6% outside Japan and exceeded the percentage of use within Japan (2.3%). Understandable translation results were obtained for 62.3% of utterances. This was a result of unsupervised adaptation for acoustic, language, and translation models using approximately 1,000,000 utterances collected from the experiment and bilingual dictionary expansion. Given that the complexity of user utterances increased with each passing day, the perplexity and number of OOVs increased and speech recognition performance degraded.

Future research will include examining even more precise unsupervised learning methods for speech recognition and translation models using large amounts of collected speech data.

References

[1] N. Bach, R. Hsiao, M. Eck, P. Charoenpornsawat, S. Vogel, T. Schultz, I. Lane, A. Waibel, and A.W. Black, "Incremental Adaptation of Speech-to-Speech Translation," Proc. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp.149–152, 2009.

[2] K. Yasuda and S. Matsuda, "7-2 Speech-to-Speech Translation System Field Experiments in All Over Japan," J. National Institute of Information and Communications Technology, vol.59, no.3/4, pp.213–223, 2012.

[3] T. Shimizu, Y. Ashikari, N. Kimura, G. Itoh, S. Matsuda, and S. Nakamura, "Evaluation of Cellular Phone Japanese-Chinese Speech Translation Experimental Service," Proc. Acoustical Society of Japan Spring Meeting, 3-Q-27, pp.269–270, 2009.

[4] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR Multilingual Speech-to-Speech Translation System," IEEE Trans. Audio, Speech, Language Process., vol.14, no.2, pp.365–376, 2006.

[5] W. Wahlster, Verbmobil: Foundations of Speech-to-Speech Translation, Berlin, Germany: Springer-Verlag, 2000.

[6] F. Metze, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, E. Pianta, J. McDonough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, K. Laskowski, and L. Levin, "The NESPOLE! Speech-to-Speech Translation System," Proc. the second international conference on Human Language Technology Research, pp.378–383, 2002.

[7] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, "Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, vol.1, pp.825–828, 2005.
[8] SLTC e-Newsletter, "DARPA's GALE Program to Get More Challenging in 2007" (Example from GALE Project): http://ewh.ieee.org/soc/sps/stc/News/NL0701/NL0701-GALE.htm
[9] S. Sakti, M. Paul, A. Finch, S. Sakai, T.T. Vu, N. Kimura, C. Hori, E. Sumita, S. Nakamura, J. Park, C. Wutiwiwatchai, B. Xu, H. Riza, K. Arora, C.M. Luong, and H. Li, "A-STAR: Toward Translating Asian Spoken Languages," Computer Speech and Language Journal, vol.27, no.2, pp.509–527, 2013.
[10] C. Hori, H. Kashioka, and E. Sumita, "Breaking Down the Language Barrior Among 23 Countires: Network-Based Speech Translation Communication Protocol Based on ITU Standards," New Breeze, vol.24, no.4, Autumn, 2012.
[11] N. Kimura, T. Shimizu, Y. Ashikari, and S. Nakamura, "A Study on Communication Interface for Multilingual Speech Translation Basis," Proc. Acoustical Society of Japan Spring Meeting, 3-Q-17, pp.249–250, 2007.
[12] K. Yasuda, E. Sumita, G. Kikui, S. Yamamoto, and M. Yanagida, "Real-Time Evaluation Architecture for MT Using Multiple Backward Translations," Proc. Recent Advances in Natural Language Processing, pp.512–522, 2003.
[13] M. Fujimoto and S. Nakamura, "A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.922–930, 2006.
[14] S. Matsuda, T. Jitsuhiro, K. Markov, and S. Nakamura, "ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.989–997, 2006.
[15] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. Speech and Audio Processing, vol.2, no.2, pp.291–298, 1994.
[16] H. Yamamoto, S. Isogai, and Y. Sagisaka, "Multi-Class Composite N-gram Language Model," Speech Communication, vol.41, no.2-3, pp.369–379, 2003.
[17] T.T. Vu, D.T. Nguyen, M.C. Luong, and J.-P. Hosom, "Vietnamese Large Vocabulary Continuous Speech Recognition," Proc. EUROSPEECH, pp.1689–1692, 2005.
[18] P. Koehn, F.J. Och, and D. Marcu, "Statistical Phrase-Based Translation," Proc. the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol.1, pp.48–54, 2003.
[19] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," Proc. the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp.177–180, 2007.
[20] G. Kikui, T. Takezawa, and S. Yamamoto, "Multilingual Corpora for Speech-to-Speech Translation Research," Proc. International Conference on Spoken Language Processing, pp.Spec3801o.2, 2004.
[21] T. Takezawa, "Building a Bilingual Travel Conversation Database for Speech Translation Research," Proc. Oriental COCOSDA Workshop, pp.17–20, 1999.
[22] H. Okuma, H. Yamamoto, and E. Sumita, "Introducing a Translation Dictionary into Phrase-Based SMT," IEICE Trans. Inf. & Syst., vol.E91-D, no.7, pp.2051–2057, 2008.
[23] Chasen: http://chasen-legacy.sourceforge.jp/
[24] Mecab: http://mecab.googlecode.com/svn/trunk/mecab/doc/
[25] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis," Proc. EUROSPEECH, vol.5, pp.2347–2350, 1999.
[26] H. Zen, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A Hidden Semi-Markov Model-Based Speech Synthesis System," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.825–834, 2007.
[27] H. Kawahara, "Speech Representation and Transformation using Adaptive Interpolation of Weighted Spectrum: VOCODER Revisited," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, vol.2, pp.1303–1306, 1997.
[28] F.K. Soong, W.-K. Lo, and S. Nakamura, "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words," Proc. SWIM, 2004.
[29] F. Wessel and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," Proc. IEEE Automatic Speech Recognition and Understanding, pp.307–310, 2001.
[30] R. Gretter and G. Riccardi, "On-line Learning of Language Models with Word Error Probability Distributions," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, vol.1, pp.557–560, 2001.
[31] R. Isotani, S. Matsuda, T. Hayashi, H. Kawai, and S. Nakamura, "Operation of a Nation-Wide Field Experiment of Speech-to-Speech Translation and Adaptation of Speech Recognition Models Using its Log Data," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J96-D, no.1, pp.209–220, 2013.

Shigeki Matsuda received his Ph.D. degree from Japan Advanced Institute of Science and Technology (JAIST) in 2003, and joined ATR Spoken Language Communication Research Laboratories as a researcher. From 2009, he joined Spoken Language Communication Laboratory of the National Institute of Information and Communications Technology (NICT) as a researcher. From 2014, he works for ATR-Trek. He is the manager of advanced technology development department. He is engaged in research on speech recognition and speech signal processing, and is a member of the Acoustic Society of Japan, Information Processing Society of Japan, and IEICE.

Teruaki Hayashi received his B.S degree from Osaka University in 1976. Between 2001–2009, he worked with the Spoken Language Communication Research Laboratories of Advanced Telecommunications Research Institute International (ATR). From 2009 to 2016, he worked at National Institute of Information and Communications Technology. From 2016, he works for ATR-Trek.

Yutaka Ashikari received his B.S. and M.S. degrees in biochemistry from Osaka University in 1982 and 1984, respectively. Between 2001–2009, he worked with the Spoken Language Communication Research Laboratories of Advanced Telecommunications Research Institute International (ATR), Japan, where he engaged in the development of speech-to-speech translation systems. Since 2009, he has been working as the Director of the System Development Office of National Institute of Information and Communications Technology (NICT), Japan, where he is engaged in the development of speech communication systems.

Yoshinori Shiga has been involved in speech technology research since 1987, at various institutions including the Tokyo University of Science, Toshiba Corporation, the University of Edinburgh, the University of Surrey, Advanced Telecommunications Research Institute International and the National Institute of Information and Communications Technology (NICT). Currently he is the leader of the speech synthesis team at NICT's Advanced Speech Technology Laboratory. He holds B.E. and M.E. degrees in Electrical Engineering from the Tokyo University of Science, and a Ph.D. degree in Speech Technology from the University of Edinburgh. He received the 2002 Information and Systems Society Excellent Paper Award and the 2014 Best Paper Award both from IEICE, and the 2015 TELECOM System Technology Award from the Telecommunications Advancement Foundation.

Hidenori Kashioka completed his Ph.D. (Engineering) program in 1993. He entered ATR the same year and later joined NICT in 2006. After serving as the Planning Manager of the General Planning Department and Director of the Spoken Language Communication Laboratory, he became the Research Executive Director of Center for Information and Neural Networks in 2013. Dr. Kashioka is primarily engaged in natural language processing and spoken language processing research. He aims for the integration of brain science and spoken language processing.

Keiji Yasuda received M.E. and Ph.D. degrees from Doshisha University in 2001 and 2004, respectively. He joined ATR Spoken Language Translation Research Laboratories in 2001, National Institute of Information and Communications Technology in 2009, and KDDI R&D Laboratories in 2013. He is currently a research engineer at KDDI Research. His research interests in include natural language processing and educational technology. He is a member of IEICE, IPSJ and ANLP.

Hideo Okuma received the B.S. degree in mathematics from Tokyo University of Science 1987. From 2003 to 2009, He was a researcher at Advanced Telecommunications Research Institute International. From 2009 to 2015, He was a researcher at National Institute of Communications Technology. He is currently a researcher at FEAT ltd. His research interests include machine translation.

Masao Uchiyama is a senior researcher of the National Institute of Information and Communications Technology, Japan. His main research field is natural language processing. He completed his doctoral dissertation at the University of Tsukuba in 1997. His current main research field is machine translation.

Eiichiro Sumita received his PhD in Engineering from Kyoto University in 1999, and a Master's and Bachelor's Degree in Computer Science from the University of Electro-Communications in 1982 and 1980, respectively. He is now an Associate Director General of the Advanced speech translation research and development promotion center (ASTREC) of National Institute of Information and Communications Technology (NICT). Before joining NICT, he worked for the Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan and for IBM Research in Tokyo, Japan. He has published 74 patents and 230 papers in the field of machine translation and e-Learning. He is a co-recipient of Minister for Internal Affairs and Communications Award of 11th annual Merit Awards for Industry-Academia-Government Collaboration in 2013, the Maejima Hisoka Prize in 2013, the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, Prizes for Science and Technology in 2010, and he was awarded the AAMT Nagao Award in 2007 and 2014.

Hisashi Kawai received B.E., M.E., and D.E. degrees in electronic engineering from The University of Tokyo, in 1984, 1986, 1989, respectively. He joined Kokusai Denshin Denwa Co. Ltd. in 1989. He worked for ATR Spoken Language Translation Research Laboratories from 2000 to 2004, where he engaged in the development of text-to-speech synthesis system. From Oct. 2004 to Mar. 2009 and from April 2012 to Sept. 2014 he worked for KDDI R&D Laboratories, where he was engaged in the research and development of speech information processing, speech quality control for telephone, speech signal processing, acoustic signal processing, and communication robots. From April 2009 to March 2012 and since Oct. 2014, he has been working for National Institute of Information and Communications Technology (NICT), where he is engaged in development of speech technology for spoken language translation. He is a member of Acoustical Society of Japan (ASJ) and IEEE.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994–2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000–2008 and Vice president of ATR in 2007–2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009–2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.