7-3 VoiceTra Field Experiments

MATSUDA Shigeki, YASUDA Keiji, and KAWAI Hisashi

We have developed a network-based speech-to-speech translation system "VoiceTra" for smart-phones that interprets users' speech into speech of foreign languages, and made it available to the public at no charge. This article briefly introduces the technologies of speech-to-speech translation and shows performance improvement obtained by using huge amount of real speech data collected by the "VoiceTra".

Keywords
Speech to speech translation, Speech recognition, Language translation, Smart-phone

1 Introduction

At the end of July 2010, the Spoken Language Communication Laboratory and the Multilingual Translation Laboratory of the Universal Communication Research Institute, NICT, released a network-based multilingual speech translation application VoiceTra (hereafter referred to as VoiceTra) as an application for Apple's smart phone iPhone, free of charge, in order to widely announce the research results of our automatic multilingual speech translation technologies and as a means of conducting field experiments for the improvement of performance using obtained data. The experiments started also for Android OS-based smart phones in April 2011. This system is mostly used to support travel conversations. For example, it is assumed to be used in communication with foreign visitors in Japan and in conversations with local people on trips abroad. In this paper we explain the configuration of VoiceTra and its speech recognition and language translation systems.

2 Multilingual speech translation application "VoiceTra"

VoiceTra is a network-based multilingual speech translation application for iPhone and Android OS-based smart phones. The start screen of VoiceTra is shown in the left panel of Fig. 1, the translation screen in the center, and the language selection screen in the right. The example shown is a translation from Japanese to English. On the upper field, the speech recognition result of the user's voice "Michini Mayoi Mashita. Eki wa Dokodesuka. (in Japanese)" is presented and the translation result "I'm lost. Where is the station?" is shown on the lower field. The Japanese text on the center field is the result of back translation from English to Japanese. The direction of the translation can be switched by tapping the arrow at the top of the screen, to translate the foreign speech of a foreigner to Japanese. The translation language can be changed by tapping the language fields, which are currently shown as "Japanese" and "English". When the language field is tapped, the right figure of Fig. 1 appears and languages can be selected as you like.

Table 1 shows the translation languages. As seen in the table, voice input using speech recognition and voice output using speech synthesis can be made for 6 languages. For a total of 21 languages, including these 6 languages, text translation can be made.

Figure 2 shows the system configuration of VoiceTra. As shown in the figure, the user's

Fig.1 *Start screen (left), translation screen (center), and language selection screen (right) of VoiceTra*

voice is transmitted through the Internet to the multilingual speech translation server. The server conducts speech recognition, language translation and speech synthesis, and each result is sent back to the user's smart phone.

Figure 3 shows the total number of accesses to VoiceTra from August 2010. As seen in the figure, the number has steadily increased since the release of the application. Till May 2012, the total number of accesses was 7.5 million. Accesses in Japanese make up 76%, those in English 19%, and those in Chinese 4%. We are currently working on the classification of the collected voice data by user attributions such as male/female, native/non-native, and others and by using situation and utilization forms.

3 Multilingual speech recognition system

To realize high-precision speech recognition, it is important to make an appropriate model to respond to users' differences, changes of speech styles, speech distortion or clipping due to background noise, and other various distortions. Research on statistical speech recognition, where a statistical model is used for speech recognition to handle these variations and distortions, has been conducted since the 1980s. VoiceTra uses a method based on statistical speech recognition. As an acoustic model, which models the time change of voices, we employed a Hidden Markov Model [1]. As a language model, which models the word order and other language information, we used an N-gram model [2]. With these models, a word sequence W^* is searched, where W^* gives the highest conditional probability $P(W|O)$ with the feature vector as time series O of an input. This is expressed by the following formula.

$$W^* = \arg\max_w P(W \mid O)$$
$$= \arg\max_w \frac{P(O \mid W)P(W)}{P(O)}$$

Here, $P(O \mid W)$ represents the acoustic model and calculates the acoustic likelihood of the feature vector sequence O with the word sequence W. $P(W)$ shows the language model and calculates the language probability of the word sequence W. arg max expresses the searching of the word sequence W^* which gives the largest probability $P(O \mid W)P(W)$. This is done by the speech recognition software. $P(O)$ in the denominator is a constant and does not have to be taken into account in

Table 1 *Translation languages*

Languages in which voice input and output are possible	Languages in which text translation can be made
Japanese, English, Chinese, Indonesian, Vietnamese, and Korean	Japanese, English, Chinese, Taiwanese, Korean, French, German, Hindi, Indonesian, Italian, Malay, Portuguese, Portuguese (Brazil), Russian, Spanish, Tagalog, Thai, Vietnamese, Arabic, Dutch, and Danish

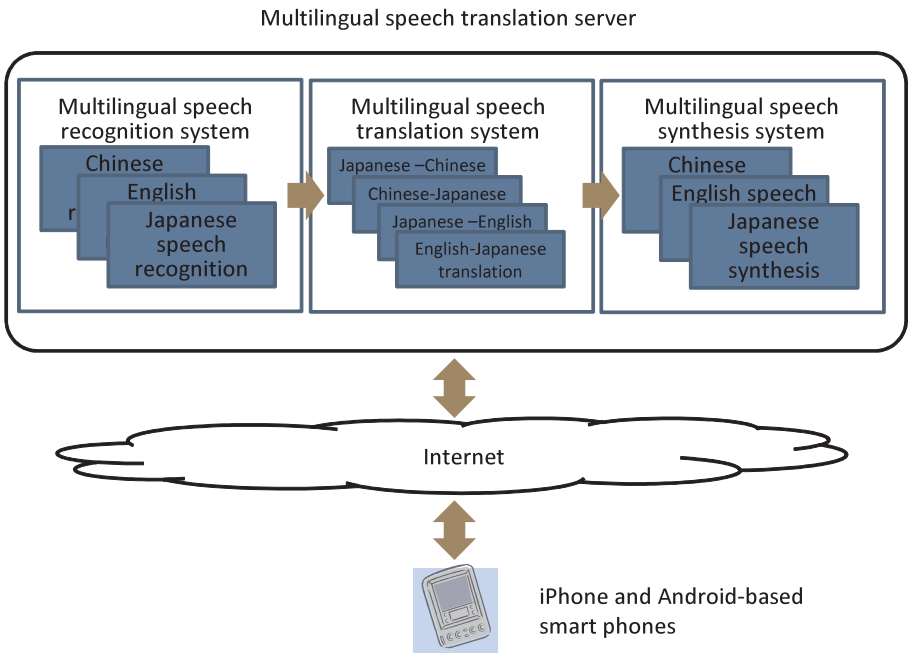

Fig.2 *System configuration of VoiceTra*

the calculation of the arg max. The model used in this statistical speech recognition is estimated from a large volume of speech and text corpora.

In the speech recognition system used when the VoiceTra service was launched, about 400-hour Japanese speech data and about 60,000 sentences were used to estimate the acoustic model. The speech voice data was collected from the voices of 4,200 adults and 300 elderly people who read out travel conversation text. The 60,000 sentences were obtained by the manual transcription of speech data collected in speech translation field experiments in 5 areas in Japan. In the field experiments conducted in these 5 areas, speech translation terminals were rented to travelers at hotels and event sites and their voices were collected from the terminals. The collected voice data contains various speech styles that are not observed in the text-reading voices.

VoiceTra is assumed to be used not only indoor but also in a noisy outdoor environment. For the improvement of the noise-robust feature, we employed a front-end process noise suppression method [3] that used the Wiener filter to collect voices. We also employed a back-end process of contaminated automobile sounds and other noises recorded on streets, at stations and at various other places for learning data to estimate the acoustic model.

The language model was estimated using 6.1 million words in the travel conversation texts and the transcribed voice data obtained in the field experiments in the five areas,

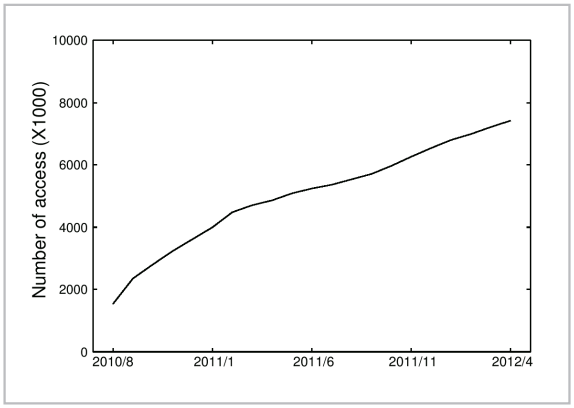

Fig.3 *Total number of accesses to VoiceTra*

which was also used for the acoustic model.

After the start of the service, the data volume has increased every day as shown in Fig. 3. By making unsupervised adaptations with this large volume of data, we attempted to improve the performance of both the acoustic and language models. Unsupervised adaptation is model adaptation without preparing transcription texts, which are usually necessary for acoustic or language models. Unsupervised adaptation uses highly reliable sentences and words, which are selected by the calculation of the reliability of each speech recognition result.

4 Language translation

The machine translation system mostly consists of statistical machine translation and two translation memories. For the statistical machine translation system, a phrase-based statistical machine translation framework [4] was used. In this framework, the probability of a word sequence (e) translated in a target language from its source word sequence (f) is calculated from the following formula.

$$p(e \mid f) = \frac{\exp\left(\sum_{i=1}^{M} \lambda_i h_i(e, f)\right)}{\sum_{e'} \exp\left(\sum_{i=1}^{M} \lambda_i h_i(e', f)\right)} \quad (1)$$

Here, e' represents a candidate translation of f. $h_i(e, f)$ is a feature function obtained from a learning corpus. There are eight feature functions, including translation probabilities (translation models) of a word or phrase from target language to source language or from source language to target language, and language models of target languages [5]. λ_i and M are a weight of each feature function and the number of the feature functions (8), respectively.

Assuming that the denominator of Equation (1) is constant, we derive a translation result $\hat{e}$ from the following Equation (2).

$$\hat{e}(e, \lambda_1^M) = \arg\max_e \sum_{i=1}^{M} \lambda_i h_i(e, f) \quad (2)$$

The basic travel expression corpus (BTEC) is mostly used as learning data. Also, the MOSES tool kit [5] and SRILM tool kit [6] are used for the learning of the translation model and language model.

Table 2 *Evaluation results of speech translation system*

System	Evaluation results			
	S	S, A	S, A, B	S, A, B, C
At start of service	24%	32%	39%	45%
After system update	33%	44%	52%	56%

5 Evaluation experiment

Table 2 shows the evaluation results of the speech translation system. In the evaluation we randomly sampled 676 sentences from the field data of VoiceTra and used them as a test set. A five-grade subjective evaluation was made by bilingual graders: S (Perfect), A (Correct), B (Fair), C (Acceptable), and D (Nonsense).

In Table 2, the performance of VoiceTra at the start of the service and after the system update is shown. In the system update, retraining of the speech recognition system and the machine translation system was made using actual data collected from VoiceTra. As shown in Table 2, the performance of the speech translation system was improved by more than 10% by using the data from VoiceTra.

6 Conclusions

We outlined the network-based multilin-

gual translation system VoiceTra, released for smart phones in August 2010 and explained the system configuration and the element technologies (speech recognition system and machine translation system) that composed the speech translation system. In the future we will examine its application to not only travel conversation but also business conversation. We will also make research and development into speech translation using track records and of an application for simultaneous interpretation.

References

1 L. R. Rabiner et al., "An Introduction to Hidden Markov Models," IEEE Transactions on Acoustic Speech, Signal Processing, Vol. 3, No. 1, pp. 4–16, 1986.

2 L. R. Bahl et al., "A maximum likelihood approach to continuous speech recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 179–190, 1983.

3 M. Fujimoto et al., "A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging," IEICE Transactions on Information and Systems, Vol. E89-D, No. 11, pp. 2783–2793, 2006.

4 P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," Proc. of HumanLanguage Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pp. 127–133, 2003.

5 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," Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180, Association for Computational Linguistics, June 2007.

6 A. Stolcke, "SRILM - an extensible language modeling toolkit," Proceedings of the International Conference on Spoken Language Processing, pp. 901–904, 2002.

(Accepted June 14, 2012)

***MATSUDA Shigeki**, Ph.D.*

Senior Researcher, Spoken Language Communication Laboratory, Universal Communication Research Institute

Signal Processing, Speech Recognition

***YASUDA Keiji**, Dr. Eng.*

Senior Researcher, Multilingual Translation Laboratory, Universal Communication Research Institute

Machine Translation, Natural Language Processing

***KAWAI Hisashi**, Dr. Eng.*

KDDI R&D Labs. Inc./ Former Executive Researcher, Spoken Language Communication Laboratory, Universal Communication Research Institute

Speech Information Processing, Speech to Speech Translation