Development of client-server speech translation system on a multi-lingual speech communication platform

Tohru Shimizu, Yutaka Ashikari, Eiichiro Sumita, Hideki Kashioka, and Satoshi Nakamura

National Institute of Information and Communications Technology,
ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan

ABSTRACT

This paper describes a client-server speech-to-speech translation system developed on a multi-lingual speech communication platform. This platform enables easy assembly of speech communication system from the corresponding software modules (e.g. speech recognition, spoken language machine-translation, speech synthesis). This client-server speech translation system is designed for use at mobile terminals. Terminals and servers are connected via a 3G public mobile phone networks, and speech translation services are available at various places with thin client. This system realizes hands-free communication and robustness for real use of speech translation in noisy environments. A microphone array and new noise suppression technique improves speech recognition performance, and a corpus-based approach enables wide coverage, robustness and portability to new languages and domains. Recent evaluation of the overall system showed that the utterance correctness of speech recognition output achieved 83%, and more than 88% of the utterances are correctly translated for Japanse-English and Japanese-Chinese.

Index Terms— speech-to-speech translation system, corpus-based, client-server

1. INTRODUCTION

Recent progress in corpus-based speech and language processing technology has made it possible to realize speech (speech-to-speech) translation in real situations. A multi-lingual speech communication platform has been developed, and a client-server speech translation system on this platform is constructed for evaluating the latest speech translation technology in real situations. All component software modules take a corpus-based statistical approach, where acoustic and language models of each module and corresponding parameters are automatically constructed (tuned) from large-scale corpora[1]. The advantages of the corpus-based approach are that it achieves wide coverage, robustness and portability to new languages and domains.

This paper presents a real-time speech translation system with microphone array processing for hands-free speech communication, new techniques for obtaining efficient models, statistical models such as HMMs (Hidden Markov Models) and Ngrams for speech recognition, statistical and example-based translation for machine translation, waveform concatenation for speech synthesis, and a result evaluation test including dialog in a real environment.

2. SPEECH-TO-SPEECH TRANSLATION SYSTEM

2.1. Multi-lingual speech translation system

The speech translation system consists of a "module manager" and component modules, i.e., automatic speech recognition (ASR), machine translation (MT), and speech synthesis (SS). The module manager has the function of controlling messages comprising speech data, recognized or translated text data, and system messages according to the predefined message flow. For example, the message flow for speech-to-speech translation is defined as follows; input speech data for each terminal are sent to the ASR,MT, SS, in this order to obtain the recognized text, translated text and its corresponding synthesized speech. Finally, these data are sent back to the terminal. To avoid degradation in the speech recognition performance caused by the loss of data packets between the terminal and the server, TCP protocol is used.

Figure 1 shows an overall configuration of the client-server speech-to-speech translation system. This system consists of several terminals and a speech translation server. The terminals and server are connected via a wireless data network. Instead of a client-server architecture shown in

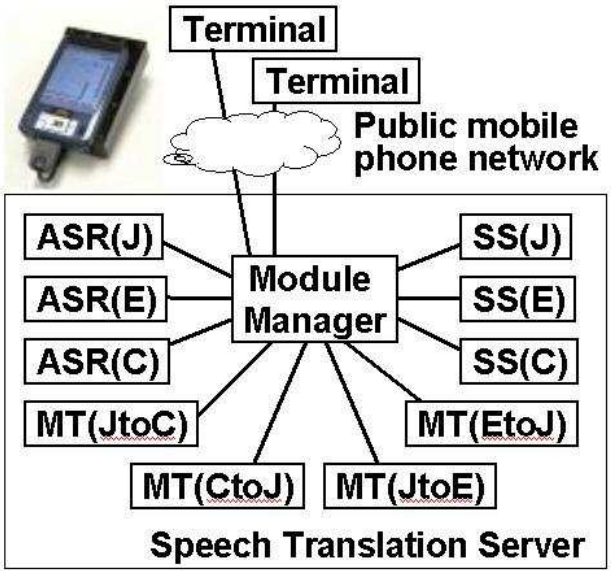

Figure 1. Configuration of the client-server speech-to-speech translation system.

figure 1, packing the entire speech translation function into one terminal is available by modifying the message flow.

2.2. Speech Translation Terminal

On each terminal, an eight-channel microphone array is mounted to realize noise robustness and hands-free speech recognition in noisy environments[2]. The input device consists of an eight-element microphone array, microphone pre-amplifiers, and an eight-channel analog/digital converter. Figure 2 shows a picture of the terminal with the microphone array. The size of the microphone array unit is W93 mm x D33 mm x H132 mm. To capture distant speech with high-quality sound, omni directional condenser microphones of type DPA 4060 are used. The microphones are arranged in a reverse L-shape with 5 microphones on the top of the PDA with sensor spacing of 2 cm and 4 microphones along the right side of the array with 4 cm spacing.

Multi-channel speech signals input from the microphone array are digitized, beam formed, and compressed in the terminal, in order to transmit speech data through a 3G public mobile phone data network. All standard sampling rates from 8 kHz to 48 kHz (8, 11.025, 16, 22.05, 32, 44.1, 48 kHz) are supported. Currently, a 16 (or 11.025) kHz sampling rate is used. The bit-depth is 16 bits.

The display has essentially three windows. The first window is for system messages, and the other windows are for recognition results and the translation result, respectively.

2.3. Component modules

2.3.1. Signal processing and speech recognition

Figure 2. Mobile multi-channel speech input device with PDA terminal. Eight microphones are arranged as reverse L-shaped array.

A robust speech recognition technology in noisy environments is an important issue for speech translation in real environments. An MMSE (Minimum Mean Square Error) estimator for log Mel-spectral energy coefficients usin g a GMM (Gaussian Mixture Model)[3], and a beam forming technique of FBF (Fixed Beam Forming) or RGSC (Robust Generalized Sidelobe Canceller)[2], are introduced for suppressing interference and noise and for attenuating reverberation.

Even when a 3G mobile phone data network is applied, the throughput of the uplink is limited. Therefore, FBF (Fixed Beam Forming) and ADPCM coding / DSR (Distributed Speech Recognition)[4, 5] is performed in the terminal. Thus, the eight-channel PCM speech data (approx. 2 Mbps for 16 kHz sampling, 16 bit quantization) is downsized to 64 kbps / 5.7kbps. When a wideband data network such as a wireless LAN (IEEE802.11b) is available, the input multi-channel speech data are transmitted to the server, and a more effective beam forming technique called RGSC (Robust Generalized Sidelobe Canceller) is applied instead of FBF.

As a technique to obtain a compact and accurate model from limited size corpora, MDL-SSS and composite multi-class N-gram are proposed for the acoustic and language modeling, respectively. MDL-SSS is an algorithm that

automatically determines the appropriate number of parameters according to the size of the training data based on the Maximum Description Length (MDL) criterion[6]. Japanese, English, and Chinese acoustic models were trained using the data of 4,200, 532, and 536 speakers, respectively. And these models were adapted to several accents, e.g., US (the United States), AUS (Australia), and BRT (Great Britain) for English. A statistical language model using the large-scale corpora (Japanese 852k sentences, English 710k sentences, Chinese 510k sentences) of travel domain corpora[7] was trained using composite multi-class Ngram[8].

Even when the acoustic and language models are well trained, environmental conditions such as speaker variability, a mismatch between the training and testing channels, or interference from environmental noise may cause recognition errors. These erroneous utterances can then be rejected by tagging them with a low confidence measure. A generalized word posterior probability (GWPP) based utterance ejection is introduced for the post processing of speech recognition[9, 10].

2.3.2 Machine translation

The translation modules are automatically constructed from large-scale corpora in the travel domain. We used TATR, a phrase-based SMT system built within the framework of feature based exponential models, and EM by exact match [11]. We employed new approaches for pre-processing (word segmentation, sentence splitting), post-processing (punctuation, capitalization) and language model adaptation[11].

The decoding process was divided into several steps: (1)For a given ASR output without case and punctuation, we used the SRI tools to insert punctuation into the output. (2)The ASR output was split according to the inserted punctuation. (3)We translated each split segment separately using TATR, and EM. (5)We recombined all segment translations to obtain the translation output. (6)We inserted punctuation and capitalization to obtain the final translation output.

2.3.3 Speech synthesis

An ATR speech synthesis engine, called XIMERA, was developed using large corpora (a 110-hour corpus of a Japanese male, a 60-hour corpus of a Japanese female, and a 20-hour corpus of a Chinese female). This corpus-based approach makes it possible to save the naturalness and personality of the speech without introducing signal processing to the speech segment[12]. XIMERA's HMM(Hidden Markov Model)-based statistical prosody model is automatically trained, so it can generate a highly natural F0 pattern[13]. In addition to that, because the cost function for segment selection is optimized based on perceptual experiments, the naturalness of the selected segments is improved[14].

3. EVALUATION OF THE SPEECH TRANSLATION SYSTEM

3.1. Speech and language corpora

We have collected three kinds of speech and language corpora such as BTEC (Basic Travel Expression Corpus), MAD (Machine Aided Dialog), and FED (Field Experiment Data)[15-18].
Especially BTEC includes parallel sentences in two languages composed of popularly used sentences in international travels. MAD is a dialog corpus collected using speech-to-speech translation system. While the amount of this corpus is relatively limited, the corpus is used for adaptation and evaluation. FED is a corpus collected in Kansai International Airport uttered by real travelers arrived at the airport.

3.2 Speech recognition system

Acoustic models are trained using 4000 speakers for Japanese, 500 speakers both for English and Chinese. The sizes of the vocabulary are around 35k in a canonical form and 50k with pronunciation variations. Recognition results are shown in the table 1 for Japanese, English and Chinese for the real time factor of 5. The real time factor is an time ratio to the length of an utterance. Although the speech recognition performance for dialog speech is worse than that for read speech, the utterance correctness excluding erroneous recognition output using GWPP[9] achieved 83% for all conditions.

Table 1. Performance of speech recognition.

	BTEC	MAD	FED
Characteristics	Read speech	Dialog speech (Office)	Dialog peech (Airport)
# of speakers	20	12	6
# of utterances	510	502	155
# of word tokens	4,035	5,682	1,108
Average length	7.9	11.3	7.1
Perplexity	18.9	23.2	36.2
%word accuracy			
Japanese	94.9	92.9	91.0
English	92.3	90.5	81.0
Chinese	90.7	78.3	76.5
% utterance correct (Japanese)			
all utterance	82.4	62.2	69.0
excluding rejected utterance	87.1	83.9	91.4

3.3 MachineTranslation

The performance of the machine translation is evaluated based a subjective scoring. The translation outputs were ranked A (perfect), B (good), C (fair), or D (nonsense) by English / Chinese native evaluators who can understand Japanese sufficiently. The accumulative score of A-C is shown in table 2.

Table 2. Performance of machine translation. (% correctly translated)

	BTEC
Japanese-to-English	92.5
Japanese-to-Chinese	88.4
English-to-Japanese	92.5
Chinese-to-Japanese	92.5

4. CONCLUSION

This paper presents the configuration and performance of a speech-to-speech translation system implemented on multi-lingual speech communication platform. This platform enables easy assembly of speech communication system from the corresponding software modules. As the terminals and server of this platform are connected via a 3G public mobile phone network, speech translation services are available at various places with thin client. Various techniques, such as a microphone array, noise suppression, and large-scale corpus based modeling for both speech recognition and machine translation realize robustness and portability for new languages and new domains by simply preparing the necessary corpora.

The experimental results evidenced effectiveness and usefulness by subjective evaluation especially with utterance rejection algorithm.

5. REFERENCES

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. The ATR multilingual speech-to-speech translation system. *IEEE Trans. on Audio, Speech, and Language Processing*, 14, No.2:365–376, 2006.

[2] W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamura. Hands-free speech recognition and communication on PDA using microphone array technology. In *Proc. of ASRU*, pages 302–307, 2005.

[3] M. Fujimoto and Y. Ariki. Combination of temporal domain SVD based speech enhancement and gmm based speech estimation for ASR in noise - evaluation on the AURORA II database and tasks. In *Proc. of Eurospeech*, pages 1781–1784, 2003.

[4] ETSI standard document. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithm. *ETSI ES 201 108*, V1.1.2, 2000.

[5] ETSI standard document. Speech processing, transmission and quality aspects(STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm. *ETSI ES 202 050*, V1.1.1, 2002.

[6] T. Jitsuhiro, T. Matsui, and S. Nakamura. Automatic generation of non-uniform context-dependent HMM topologies based on the MDL criterion. In *Proc. of Eurospeech*, pages 2721–2724, 2003.

[7] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In *Proc. Of Eurospeech*, pages 381–384, 2003.

[8] H. Yamamoto, S. Isogai, and Y. Sagisaka. Multi-class composite N-gram language model. *Speech Communication*, 41:369–379, 2003.

[9] Frank K. Soong, Wai-Kit Lo, and Satoshi Nakamura. Optimal acoustic and language model weigths for minimizing word verification errors. In *Proc. of ICSLP*, pages 441–444, 2004.

[10] Wai Kit Lo and Frank K. Soong. Generalized posterior probability for minimum error verification of recognized sentences. In *Proc. of ICASSP*, pages 85–88, 2005.

[11] Ruiqiang Zhang, Hirofumi Yamamoto, Michael Paul, Hideo Okuma, Keiji Yasuda, Yves Lepage, Etienne Denoual, Daichi Mochihashi, Andrew Finch,Eiichiro Sumita, "The NiCT-ATR Statistical Machine Translation System for the IWSLT 2006 Evaluation," submitted to IWSLT, 2006.

[12] H. Kawai, T. Toda, J. Ni, and M. Tsuzaki. XIMERA: A new TTS from ATR based on corpus-based technologies. In *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. of ICASSP*, pages 1215–1218, 2000.

[14] T. Toda, H. Kawai, and M. Tsuzaki. Optimizing sub-cost functions for segment selection based on perceptual evaluation in concatenative speech synthesis. In *Proc. of ICASSP*, pages 657–660, 2004.

[15] T. Takezawa and G. Kikui. Collecting machine –translation-aided bilingual dialogs for corpus-based speech translation. In *Proc. of Eurospeech*, pages 2757–2760, 2003.

[16] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In *Proc. Of Eurospeech*, pages 381–384, 2003.

[17] T. Takezawa and G. Kikui. A comparative study on human communication behaviors and linguistic characteristics for speech-to-speech translation. In *Proc. of LREC*, pages 1589–1592, 2004.

[18] G. Kikui, T. Takezawa, M. Mizushima, S. Yamamoto, Y. Sasaki, H. Kawai, and S. Nakamura. Monitor experiments of ATR speech-to-speech translation system. In *Proc. of Autumn Meeting of the Acoustical Society of Japan*, pages 1–7–10, 2005, in Japanese.