SPEECH-TO-SPEECH REAL-TIME TRANSLATION

P.E. Bukreev, D.A. Morel Morel

Russia, Belgorod State National Research University

The given article considers the principles and technologies of Speech-to-Speech Real-Time Translation, covers general and specific problems this IT area faces and overviews its main achievements.

Speech-to-Speech (S2S) Real-Time Translation is the on-the-fly machine translation using special software and hardware to translate speech from one language to another. It is one of the state-of-the-art R & D areas related to the construction of machine translation systems. This technology allows native speakers of different languages to communicate with each other in real time. Thus, it is of great importance for humanity, as it involves the development of science, business, culture, education system.

This process is as follows: a speaker of language A speaks into a microphone and the speech recognition module recognizes the utterance. Then it compares the input with a phonological model, consisting of a large corpus of speech data. The input is then converted into a string of words, using dictionary and grammar of language A. The machine translation module then translates this string. Early systems replaced every word with a corresponding word. The generated translation utterance is sent to the speech synthesis module, which estimates the pronunciation and intonation matching the string of words based on a corpus of speech data in language B. Data that correspond to the phrase are selected, combined and displayed in the form necessary for a speaker of language B.

Thus, the process of S2S Real-Time Translation typically integrates the following three software technologies [Lazzary 2008: 49]:

1) Automatic speech recognition (ASR),
2) Machine-Assisted Translation (MT),
3) Text-to-speech synthesis (TTS).

Automatic speech recognition is the process of converting a voice signal into digital data (for example text data). As early as 1952 the first device for automatic speech recognition appeared, it could recognize numerals pronounced by humans. Commercial speech recognition software has appeared in the early nineties. Translation reliability of these programs is not very high, but over the years, it has gradually improved. The next step in speech recognition technology is the development of so-called Silent Speech Interfaces (SSI). Such systems of speech recognition are based upon getting and processing speech signals at an early stage of articulation. This stage of speech recognition development is caused by two essential disadvantages of modern speech recognition systems: excessive sensitivity to noise and, moreover, the need for a clear and accurate speech when handling such a system. Approach based on SSI consists in using new sensors not subject to the influence of noises mixed to the processed acoustic signals [Freitas et al. 2011].

The stages of speech recognition are the following.

1. Processing of speech begins with evaluating speech signal quality. At this stage the levels of noise and distortion are defined.

2. The result of evaluation is supplied to the acoustic adaptation module that controls the speech parameters calculation module required for recognition.

3. Signal plots containing speech are selected, and speech parameter evaluation occurs. Phonetic and prosodic probabilistic characteristics for syntactic, semantic and pragmatic analysis are selected (information about parts of speech, word form and statistical relationships between words is evaluated).

4. Then the speech characteristics come into the main block of the recognition system, the decoder. This component matches the input speech stream with the information stored in the acoustic and language models, and determines the most likely sequence of words that comes as the final result of recognition.

Machine-Assisted Translation is the process of translating texts from one natural language to another using a special computer program. Also it refers to the direction of scientific research related to the construction of such systems [Kulagina 1991: 5]. There are two fundamentally different approaches to the construction of algorithms for machine translation: rule-based and statistical (or statistical-based) [Islamov, Fomin 2013].

Rule-Based Machine Translation is a general term that denotes machine translation systems based on linguistic information about source and target languages. They consist of bilingual vocabularies and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. Such an approach to machine translation is also called classic. On the basis of these data an incoming text is converted into a translated text sequentially, sentence by sentence. These systems are contrasted with machine translation systems, which are based on examples. The principle of such systems operating is the connection between the structure of input and output sentences. The approach is traditional and is used by most developers of machine translation systems (PROMPT in Russia, SYSTRAN in France, Linguatec in Germany, etc.).

Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. Statistical machine translation has the property of "self-training". The more language pairs are available and the more accurately they correspond to each other, the better result of statistical machine translation is. Statistical machine translation systems are opposed to machine translation based on the rules of Rule-Based Machine Translation (RBMT) and examples of Example-Based Machine Translation (EBMT).

The first ideas of statistical machine translation were published by Warren Weaver in 1949. The "Second Wave" had its beginning in the 1990s within IBM fold. The "Third Wave" corresponds to activities of Google, Microsoft, Language Weaver, Yandex, and also to a new service from ABBYY [Grashchenko et al. 2011: 279]. Text-to-speech synthesis is primarily called everything connected with the artificial production of human speech. Speech synthesizer is a structure in software or

hardware capable of translating text / images into speech [Lobanov, Tsirul'nik 2008: 316].

The quality of a speech synthesizer is evaluated by its similarity to the human voice and by its ability to be understood. The simplest synthesized speech can be created by combining parts of recorded speech, which then will be stored in a database.

When overviewing the history of S2S practical implementation we should mention two companies' contribution. NEC Corporation went a long way from a concept exhibit at the ITU Telecom World in 1983 to a mobile device with on-board Japanese-to-English speech translation in 2006. In 2001 Robert Palmquist, having released an English-Spanish large vocabulary system in 1997, founded SpeechGear, which has broad patents covering of speech translation systems. Having started in 2003 with the world's first commercial mobile device with on-board Japanese-to-English speech translation, in 2009 this company released a version providing instant translation of conversations between English and approximately 35 other languages.

As for the latest designs of S2S the following ones should be mentioned.
During Wireless Japan 2011 exhibition NTT DoCoMo, Japan mobile operator, presented the service of automatic mobile translation of conversation. One of employees who were in Yokosuka has been reading a newspaper in Japanese while visitors in Tokyo were listening to simultaneous translation in English. In test mode, the operator set up the service in November 2011 [Shafir 2011].

Ryuji Yamada, the president of NTT DoCoMo, announced that his team is working on the phone which will translate the conversation like in a sci-fi movie. Even today, there is impressive speed of operation of the new service, nevertheless solving the problem of the accuracy of the translation is still in store for them. The Japanese managed to combine technologies of translation, recognition and speech synthesis in the new design [Shinya, Kosuke 2012].

The company Google is not far behind the Japanese, and has recently made an experiment with implementing voice recognition in the online translator Google Translate. Meanwhile this function works only in text mode of Google Chrome web-browser and only translates English phrases.

Defense agency DARPA (USA) is working on the same designs. According to reliable sources, the system translates colloquialisms, emails and text messages in instant messengers. A simultaneous translator of lectures was developed in Germany in 2012. Researchers from Karlsruhe Institute of Technology created a program translating oral lectures by German teachers into English, the translation with subtitles appearing on the screen during the lecture. This is the result of two decades of scientific researches of scientists led by Professor Alex Waibel. The program is still in test mode. The system recognizes the German language, and a translation into English appears on the big screen or on students' tablets and smartphones. There is no need to install any applications because the translated text is sent over the Internet. The translation is not always ideal, but English-speaking students with limited knowledge of German can understand their German lecturers better. There is an intention to adapt the program to a wider range of subjects, since the system understands only lecturers from faculty of engineering and computer science. It is

promised to add a few languages. The researchers hope that the introduction of such software in education will increase the inflow of foreign students. The English language is more common than German. It is obvious that the majority of foreign students will choose educational institutions with better support of the English language. It will increase the competitiveness of German universities in the international education market [Roe 2012].

Microsoft presented an automatic, almost simultaneous voice translator from English to Putonghua in October 2012. Self-training system is based on artificial neural networks (Deep Neural Networks) and reduces the misunderstanding to each seventh-eighth word. However, the greatest achievement is the generation of speech with preservation of the speaker's voice modulations [Rick 2012].

Considering the problems the Speech-to-Speech Real-Time Translation faces we can ascertain that they are mostly shared with the machine translation in general, which quality of depends on the subject and style of the original text, as well as of grammatical, syntactic and lexical cognation of source / target languages. The more formalized is style of the original document, the greater quality of translation you can expect when using any machine translation systems. The application of machine translation without setting on the theme (or deliberately wrong setting) leads to incorrect results. This is connected with the fact that the program does not recognize the context of phrases and terms and translates them literally; moreover, it does not distinguish proper names from ordinary words. The situation worsens in case of speakers belonging to different linguistic groups. For example, English language belongs to the Germanic group of Indo-European languages, while Chinese belongs to the Sino-Tibetan language superfamily. The differences between them are great and it is difficult to make a correct translation, because the same word can have two or more different variants of meaning within the translation into another language. For these reasons, the percentage of translation errors is still high in case of languages distant from each other unlike related languages, for example, Russian and Ukrainian.

In addition to general problems of machine translation synchronous S2S technology faces challenges of speech recognition and text-to-speech synthesis.
Thus, main problems of speech recognition are spontaneous speech followed by speech agrammatisms and speech "garbage", the embolophrasia, the presence of acoustic noises and distortions including changing ones, and the presence of speech interferences. The recognition of continuous spontaneous speech is the ultimate purpose of all efforts for speech recognition. The main problem of the text-to-speech synthesis is robot-like sounding of generated voice; furthermore, such an artificial voice is sometimes difficult to comprehend.

Thus, S2S technology has its own specific problems, such as incoherence, less restrictive grammar, and obscure word boundaries of spoken language, the necessity of error corrections in speech recognition.

However, the simultaneous S2S translation has advantages over the machine translation; for example, it deals with less complex structure as well as less lexicon of spoken language.

To sum up we should point out that with increasing power of hardware devices we can expect the appearance of the machine translators that will have fewer errors in

translation, which presents to be the main problem of all electronic speech translators. Nevertheless, there are already translators capable of translating within the given subject framework with minimal errors, for example, the German Karlsruhe Institute of Technology, where the lectures for students are translated on-the-fly in form of displayable subtitles. In addition, in 2012 Microsoft developed a translator with a self-training system that provides with better translation and reproduction of the speaker's voice. There are also systems which allow the native speakers of different languages to communicate by telephone and still understand each other, as in the case of Japanese mobile operator NTT DoCoMo.

The improvements in technology will lead to automation in the field of translation and to reduction in number of vacancies for interpreters. For example, the company Skype provides simultaneous interpreter service for a fee and it is not surprising if this service becomes cheaper in the nearest future, the job being done by a program. Demand for this technology is getting higher and very soon people from different countries would understand each other regardless of language barriers.

REFERENCES:

1. Baard, E. (2003). Device: Arabic In, English Out [Electronic Resource]. –URL: http://www.wired.com/gadgets/miscellaneous/news/2003/03/58150.
2. Freitas, J., Teixeira, A., Dias, M., Bastos, C. (2011). Towards a Multimodal Silent Speech Interface for European Portuguese // Speech Technologies / Prof. Ivo Ipsic (Ed.). – InTech; University of Rijeka. – Pp. 125-150.
3. Grashchenko, L.A., Klyshinskii, E.S., Tumkovskii, S.R., Usmanov, Z.D. (2011). Konceptual'naya model' sistemy russko-tadzhikskogo mashinnogo perevoda // Doklady Akademii nauk Respubliki Tadzhikistan, 54/4. – Pp. 279-285.
4. Islamov, R.S., Fomin, A.G. (2013). Analiz sovremennyh sistem mashinnogo perevoda tipa SMT i RBMT // Filologicheskie nauki. Voprosy teorii i praktiki, 3(21)/1. – Pp. 69-73.
5. Kulagina, O.S. (1991). O sovremennom sostoyanii mashinnogo perevoda // Matematicheskie voprosy kibernetiki. – M.: Nauka. – Issue 3. – Pp. 5-50.
6. Lazzary, G. (2008). Perevodcheskie tehnologii dlya Evropy. – M.: MCBS. – 64 p.
7. Lobanov, B.M., Tsirul'nik, L.I. (2008). Komp'yuternyi sintez i klonirovanie rechi. – Minsk: Belorusskaya Nauka. – 316 p.
8. Rick, R. (2012). Speech Recognition Breakthrough for the Spoken, Translated Word [Electronic Resource]. – URL: http://research.microsoft.com/apps/video/dl.aspx?id=175450.
9. Roe, L. (2012). Simultaneous Translation: University without Language Barriers [Electronic resource]. – URL: http://www.kit.edu/visit/pi_2012_10978.php.
10. Shafir, S. (2011). NTT DoCoMo exhibits on-the-fly speech translation, lets both parties just talk [Electronic resource]. – URL: http://www.engadget.com/2011/05/30/ntt-docomo-exhibits-on-the-fly-speech-translation-lets-both-par/.
11. Shinya, I. Kosuke, T. (2012). Speech Recognition Technology and Applications for Improving Terminal Functionality and Service Usability [Electronic resource]. – URL: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical_journal/bn/vol13_4/vol13_4_079en.pdf.