Technology | DOI:10.1145/2582111

Paul Hyman

Speech-To-Speech Translations Stutter, But Researchers See Mellifluous Future

The practical need for accurate instant or simultaneous machine translations continues to grow as applications multiply.

WHILE COMPUTER SCIENTISTS have yet to build a working "universal translator" such as the one first described in the 1945 science-fiction novella "First Contact" and later employed by the crew of the Starship Enterprise on "Star Trek," the hurdles to creating one are being cleared. That is because the practical need for instant or simultaneous speech-to-speech translation is increasingly important in a number of applications.

Take, for example, the hypergrowth of social networking and Skype chats that demand bidirectional, reliable, immediate translations. Similarly, when natural disasters strike, the problem of aid workers struggling to communicate with the stricken who often speak other languages can become overwhelming.

When tourists travel to remote areas; when businesses want to speed commerce; when doctors who speak only one language need to talk with patients who only speak another; when immigration control is unable to conduct interviews because of a language barrier; all of these are good examples of why instant or simultaneous speech-to-speech translation has become more than just a good idea.

Indeed, in the European Union, where there are 24 official languages (up from 11 in 2004) and where the European Commission employs approximately 3,000 staff translators and interpreters, the cost of maintaining the EU's policy of multilingualism was 1.1 billion euros ($1.48 billion) last year.

Carnegie Mellon computer science professor Alexander Waibel (far right), one of the developers of speech translation app Jibbigo, looks on as a U.S. Marine uses the app to communicate with a Thai native.

U.S. MARINE CORPS PHOTO BY LANCE CPL. KRIS B. DABERKOE

"Simultaneous interpreting by machine, once further developed, could facilitate communication in certain contexts of simple communication where a professional human interpreter is not available," says Susanne Altenberg, head of the Unit for Multilingualism Support at the European Parliament in Brussels. "Some aspects of speech-to-speech translation could also assist human interpreters in their work."

However, the difficulty with machine interpretation that goes out live, immediately, and ensures rapid oral communication, says Altenberg, is that such revision is not possible as it is with translated texts, and current methods are not yet sufficiently reliable.

While Altenberg says it is difficult to predict how long it will take to develop a fully reliable speech-to-speech translation capability, "we follow developments with interest. Simultaneous interpreting is a very complex task involving almost all human cognitive and emotional capabilities. It is quite a challenge for computers to imitate this, especially with 552 language combinations [in the EU] and in a highly political context such as ours."

The single highest hurdle to a fully reliable speech-to-speech translation technology is that language is inherently ambiguous, says Jaime Carbonell, director of the Language Technologies Institute at Carnegie Mellon University's School of Computer Science. For instance, in English, the word "line" can mean a geometrical line, a queue where people stand in front of one another, a rope (as in "tangled up in the lines"), a railroad line, an actor's line in a script, and so on. We may not think of "line" as having multiple meanings but, for translation purposes, there are actually 16 of them.

"Traditionally, the most difficult problem—and it remains so—is that you need to use context to pick the right meaning, to resolve the ambiguity so that the correct words and phrases are chosen in the target language," Carbonell explains.

Yet pinpointing which solution is currently the 'state of the art' depends on whom you ask.

At IBM's Thomas J. Watson Research Center, Salim Roukos, senior manager of Multilingual NLP Technologies and CTO Translation Technologies, describes the speech-to-speech translation system on which he and his team are working as a budding technology consisting of three parts: a speech recognition system that converts the audio of the speaker's source language into written text; a text-to-text system to translate the text into the target language, and a text-to-speech system to synthesize audio from the written target language. All three components need to behave really well, individually and together, to achieve a good result.

However, there have always been two major difficulties with such systems, says Roukos.

The first involves speech recognition. "When I don't articulate, when I speak quickly or drop words, if I have an accent, that makes speech recognition harder and the machine does not do as well," he explains. "The error rate can increase from a few percent to tens of percents."

Yet that is no longer the greatest challenge, he says. In a recent evaluation of his team's work, Roukos says speech recognition has improved by about 40%, compared to a year ago, in terms of the reduction of word errors. "We still have a ways to go," he says, "but we have gotten to that point in our work, which is based on what we call convolutional neural networks, where speech recognition is not the toughest problem."

"When I don't articulate, when I speak quickly or drop words, if I have an accent, that makes speech recognition harder and the machine does not do as well."

The more recent challenge is tackling out-of-vocabulary (OOV) words. Because languages have dialects and speakers frequently use slang, a system's translation from the source language to the target language is often inaccurate. That is why Roukos and his team introduced the concept of a dialogue manager on both sides of the conversation. For example, if the speaker says a word the dialogue manager does not recognize, it will play to the speaker the audio that corresponds to that word, and then ask whether the word is a name; if it is not a name, it will request a synonym, a paraphrase or, in extreme cases, a spelling of the word.

The most recent metrics show that, about 80% of the time, the system is able to detect when it fails to recognize the input and then interacts with the user. Roukos expects to be able to improve that to 90%–95% in the next few years.

"What we are doing now is using machine-mediated speech-to-speech translation, in which the machine is taking an active role in helping the communication across the two languages," says Roukos. "This is brand-new, the state of the art, but a work in progress."

Carnegie Mellon's Carbonell identifies the commercialization of speech-to-speech translation as one of the most recent developments in the field. For instance, in August 2013, Face-

ACM Member News

ZYDA LEADS USC CS GAMES PROGRAM TO THE TOP

"I make change and I make my next position for the job that needs me," said Michael Zyda, founding director of the University of Southern California's GamePipe Laboratory, and a professor of engineering practice in USC's department of computer science.

In 2004, at age 50, Zyda, an ACM Distinguished Speaker, opted to reinvent himself by leaving his posts as computer science professor and founding director of the Modeling, Virtual Environments and Simulation (MOVES) Institute at the Naval Postgraduate School in Monterey, CA, where he served as principal investigator and development director of the "America's Army" PC game. He went to USC to launch the Joint Games Program, now called USC Games; within five years, he had made it the world's top computer science games program.

USC Games now has 80 engineers (including 30 interactive game designers, as well as 150 artists from outside art schools) who build approximately seven games a year. At the end of every fall and spring semester, it hosts USC Games Demo Day, attracting hundreds of top gaming industry professionals.

Zyda and his students designed Black Ops 1 and Black Ops 2; Grand Theft Auto 5 ("it made $1 billion in three days," he recalls); Modern Warfare 3, and Farmville. "You don't make a Triple A game title in America without a USC Games alumnus," Zyda said, adding, "over 90% of our students have jobs before graduation."

Zyda is a triple-threat career as a computer science professor, game designer (he co-holds the patent on the Nintendo Wii U console's nine-axis sensor), and as an expert witness in gaming industry patent litigation cases.

What does he do for fun? "I swim 2,000 meters freestyle in the USC pool every day for 40 minutes, about 27 miles a month," Zyda laughed.

—Laura DiDio

book acquired the team and technology of Pittsburgh-based Mobile Technologies, a speech recognition and machine translation startup that had spun off from Carnegie Mellon and which developed the app Jibbigo. The app allows users to select from more than 25 languages, record a voice in that language, then have a translation displayed on screen and read aloud in a language of your choosing.

"Speaking of commercialization, Google, with its Google Translate, does an excellent job of using context to determine which is the speaker's intended meaning," says Carbonell. "That is because Google has access to so much more data than do others, so it can get better statistics to determine, in the context of a group of words or phrases, what is the most likely meaning so that it can translate accordingly. That is Google Translate's strength."

Indeed, the Google Translate Android app—which allows users to speak into a phone, translates the words spoken into a different language, then allows a second user to respond in their own language—began 10 years ago as a third-party software product that translated just eight major languages. Today, the app can translate 72 different languages—from Afrikaans to Yiddish—and processes over a billion translations daily. That allows Google to gather enough data to create dictionaries automatically by learning from that data.

Meanwhile, Google is researching new ways to resolve the translation problem so it might learn more effectively.

"What we have done is to simplify the translation method," says Franz Och, who heads up the Google Translate team. "Instead of the typical method of feeding whole translated documents into the learning process, we are able to seed our vector system with just a little bit of information—about 5,000 words from Google Translate for which translations are known—and the system can then find parallels between words in different languages in documents that have not yet been translated. Basically, we have made the move from parallel texts, which is what we call texts for which we have specific translation, to just comparable text … and then use that to learn translation information."

"We are closer than ever to translation that is so quick and natural it feels like human translation."

It is important to note, says Och, that the vector system is not a translation system in itself. "It is just interesting in that it can find these parallels without our feeding it documents that have already been translated," he adds. "So there is some potential here for this vector approach to help with our existing machine translation system."

During the past three years, researchers at Microsoft Research (MSR) report having dramatically improved the potential of real-time, speaker-independent, automatic speech recognition by using deep neural networks (DNNs).

At Interspeech 2011, MSR demonstrated how to apply DNNs to large-vocabulary speech recognition and reduce the word error rate for speech by over 30% compared to previous methods, recalls Frank Seide, principal researcher and research manager at MSR's Speech Group. "This means that rather than having one word in four or five incorrect, now the error rate is one word in seven or eight," he says. "While still far from perfect, this is the most dramatic change in accuracy since 1979 and, as we add more data to the training, we believe we will get even better results."

Seide's team also found using DNNs helps recognition engines deal with differences in voices and accents and the conditions under which the speech was captured, like microphone type and background noise.

"We also determined that our work in DNNs can learn across languages," Seide adds. "In other words, example data of one language can help to improve accuracy for another language.

This is very important since speech recognizers, as part of their 'training,' must be exposed to extremely large amounts of example speech data—on the order of thousands of hours of speech—which has to be painstakingly transcribed down to the last stutter. These significant improvements have aided in the progression of new recognition scenarios such as the possibility of broad-scale speech-to-speech translation."

Microsoft recently applied MSR's DNN technology to the company's Bing Voice Search app for Windows Phone, providing a 12% improvement in word error rate overall compared to the previous Bing system.

Going forward, what are the next steps for researchers working on speech-to-speech translation?

At MSR, the goal is to improve recognition and translation of language as people use it. "People don't speak in the same way they write," says Chris Quirk, senior researcher at MSR's Natural Language Processing Group. "They do not even write like they used to; look at social media sites such as Facebook and Twitter. Being the translation service for Facebook has given us a unique view on this rapid change, driving us to broaden our systems toward the language of today and tomorrow. Clearly, we have to find new and different data sources."

At IBM, Roukos' team will concentrate on improving their system's ability to detect "low-confidence regions" where the system does not recognize certain words and/or phrases, does not know how to translate them, or is not sure it translated them correctly. "The ability of the system to detect when it does not know what the input is, and therefore is interacting with the user to clarify or paraphrase, is our core focus."

Meanwhile, at Google, Och is reluctant to predict how long it will be before a close-to-foolproof method exists for simultaneous or instant language translation. "If you asked that of people in AI research any time since the 1950s, the answer would be 'in about five years,' so that is what I will say is, in about five years. But it is true we really are closer than ever to translation that is so quick and natural it feels like real human translation.

"For some language pairs—like Spanish to English—we are already pretty close; people judge our translations to be sometimes as good as, or better than, a human translation. Of course, that is partly because human translations are not always that great. Ideally, machine translations would be even more consistently high-quality; at least, that is our goal."

Further Reading

"Simultaneous Translation By Machine," a video posted June, 2012 by KITinformatik at http://www.youtube.com/watch?v=q2amqJmmDm4

"Exploiting Similarities Among Languages For Machine Translation," Tomas Mikolov, Quoc V. Le, and Ilya Sutskever, September 2013, the Cornell University Library, http://arxiv.org/abs/1309.4168

"Learning The Meaning Behind Words," a blog by Tomas Mikolov, Ilya Sutskever, and Quoc Le, published August 2013 by Google, at http://google-opensource.blogspot.com/2013/08/learning-meaning-behind-words.html

"Breaking Down The Language Barrier – Six Years In," a blog by Franz Och, published April, 2012 by Google, at http://googleblog.blogspot.com/2012/04/breaking-down-language-barriersix-years.html

E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Déchelotte, M. Federico, M. Kolss, Y. S. Lee, J. B. Mariño, M. Paulik, et al.,
"System Combination For Machine Translation Of Spoken And Written Language," September 2008, IEEE, http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4599393&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4599393

"Speech Recognition Breakthrough for the Spoken, Translated Word," a video posted Nov. 7, 2012 by Microsoft Research, at http://www.youtube.com/watch?feature=player_embedded&v=Nu-nlQqFCKg

"Speech Recognition Leaps Forward," an article by Janie Chang, published August, 2011 by Microsoft Research, at http://research.microsoft.com/en-us/news/features/speechrecognition-082911.aspx

Google Translate, an app developed by Google, updated Nov., 2013, at https://play.google.com/store/apps/details?id=com.google.android.apps.translate

Bing Translator, an app developed by Microsoft Research, updated Dec., 2013, at http://www.windowsphone.com/en-us/store/app/translator/2cb7cda1-17d8-df11-a844-00237de2db9e

Paul Hyman is a science and technology writer based in Great Neck, NY.

Education

ACM Report Urges Expansion of CS Education

A new report from ACM found few states positioned to provide students with the computer science (CS) education required for rewarding careers and to ensure the needs of the future workforce are met.

The report, *Rebooting the Pathway to Success: Preparing Students for Computing Workforce Needs in the United States*, urges state education and business leaders and public policy officials to work together to develop comprehensive CS education and workforce development plans. The report provides recommendations to help these leaders create pathways that will expose all K–12 students to computer science, provide expanded access to more rigorous CS courses, offer increased opportunities for students to pursue post-secondary degrees, and align education pathways with computing careers.

"By 2020, one of every two jobs in science, technology, engineering, and mathematics (STEM) will be in computing," said Bobby Schnabel, chair of ACM's Education Policy Committee. "This concentration of computing positions in STEM makes it imperative for K–12 students in academic and career technical education programs to gain more opportunities to learn computer science."

The report calls on colleges and universities to play a role in expanding opportunities for computer science education by recognizing rigorous computer science courses in their admissions processes. Higher education institutions also can reduce barriers to degree completion by adopting systemwide agreements that allow students to transfer course credits to fulfill their computing degrees efficiently.

The full report is available at pathways.acm.org.