https://doi.org/10.48009/1_iis_2010_129-133

AUTOMATIC INTERPRETATION OF ENGLISH SPEECH

Milam Aiken, University of Mississippi, maiken@bus.olemiss.edu
Lakisha L. Simmons, University of Mississippi, lsimmons@olemiss.edu
Shilpa Balan, University of Mississippi, sbalan@bus.olemiss.edu

===

ABSTRACT

Automatic interpretation of human speech into different languages is difficult as it involves problems of speech recognition and synthesis as well as machine translation. Although several hand-held devices have been developed to provide pre-recorded spoken phrases, only a few are capable of uttering phrases with unrestricted dialog, and these are often limited to a few languages. This paper describes a new, portable prototype incorporating readily available free or commercial software that interprets between English and 20 different languages. A test of the system with German and Spanish shows relatively accurate results.

Keywords: Speech recognition, Machine translation, Speech synthesis, interpretation

INTRODUCTION

Human interpreters can be expensive or not available, especially for translations of relatively uncommon languages in remote locations. When interpretation is not possible, travelers might resort to a common language such as English to communicate with people from other countries [8, 9], even though non-native speakers might find using the language difficult [6, 16], and they might not be as confident and assertive when speaking a second language [10].

Telephone interpretation (TI), in which a person speaks into the receiver using one language and hands the device to another who hears a human interpreter provide the equivalent message in another tongue, allows travelers access to interpretation among hundreds of languages within seconds and at a relatively low cost, but this still could be inconvenient. In addition, it might not be appropriate for a private, sensitive conversation. To help address this language barrier, speech-to-speech machine interpretation (MI) or computer-assisted interpretation (CAI) systems are being used around the world, but primarily by the military [28]. For example, coalition forces in Iraq and Afghanistan are now using mobile devices that provide multilingual, distributed communication.

This paper reviews the current state of MI technology and describes a new system using off-the-shelf hardware and software that has the potential for providing greatly expanded mobile interpretation services. A study of the system with three languages shows that relatively accurate automated interpretation is possible.

MACHINE INTERPRETATION SYSTEMS

Research into unrestricted, automatic, speech to-speech interpretation has been ongoing for at least 15 years. In an early study [12], researchers developed ASURA, a system that could recognize spoken Japanese and output synthesized speech into equivalent English or German phrases with high accuracy and fairly good efficiency. Other research investigated interpretation between English and French [21] and Waibel [24] demonstrated the ability to automatically interpret among the English, German, Japanese, Spanish, and Korean languages.

A few years later, *Sync/Trans* was developed to provide interpretation of spontaneous speech between English and Japanese [11], and NEC created a system for interpretation of speech limited to travel situations in either Japanese or English [26]. The DIPLOMAT rapid-deployment speech translation system supported Serbo-Croatian and English [5], and the EUTRANS system supported Italian, English, and Spanish [19]. Another system using Japanese and English achieved 85-96% for speech recognition accuracy and 83-87% for comprehension [17].

Tests of an MI system called *TRIM* (trans-lingual instant messaging system interface) showed the following results: Chinese to Japanese (25%), Korean to Japanese (31%), English to Japanese (56%), and English to Chinese (67%) [15]. Another system provided speech-to-speech interpretation between English travel-related conversations and Japanese or Chinese [13]. The interpretation quality was considered high, e.g., at the level of a person having a Test of English for International Communication (TOEIC) score of 750 out of the perfect score of 990. One system limited to phrases about the weather enabled English speech to be converted to Mandarin Chinese text. Results of 695 spontaneous utterances showed 89.9% of the final text was correct, 6.1% incorrect, and 4.0% rejected [25].

MASTOR (Multilingual Automatic Speech-to-Speech Translator) translates unconstrained, free-form speech between English and Mandarin Chinese and uses back translation to judge the resulting [20]. Tests of the product using 237 spoken English utterances from 3 speakers and 77 spoken Chinese utterances from 2 speakers showed an average word error rate of about 10% for English and about 5% for Chinese [29], but the resulting translations were generally poor: The desktop BLEU score [18] for English to Chinese was only 35 out of 100 maximum, and the Chinese to English score was 29. The Ipaq BLEU score for English to Chinese was 31, and the Chinese to English score was 28.

Other research resulted in *Verbmobil*, a product that provides interpretation among German, English, and Japanese. Installed on mobile phones, people can use this system to support face-to-face or geographically-dispersed conversations. Generally restricted to three business-oriented domains, tests have shown an average processing time of four times the input signal duration, a word recognition rate of more than 75% for spontaneous speech, 80% comprehension of the resulting speech, and a 90% success rate for dialog tasks [23].

Work is currently underway at DARPA to develop a universal translator (IraqComm) that easily fits in a bag the size of a woman's purse, and tests in Iraq between English and Arabic speech show about 70% or 80% accuracy [7, 22].

A NEW, GOOGLE-BASED MACHINE INTERPRETER

Prior research of constrained and unconstrained MI has resulted in systems that are restricted in the number of languages supported, and resulting accuracies have varied between 25% and 90%. We have developed a new prototype desktop or laptop-based system that uses a commercial product, *Dragon Systems Naturally Speaking 6.0 (DS),* for speech recognition, a free Web-based service, *Google Translate (GT),* for machine translation, and a free, downloadable Active-X control, *MicroSoft Agent,* for text-to-speech (TTS) vocalization. The resulting system, integrated using Microsoft's *Visual Studio.NET 2008,* provides support for up to 20 different languages, with varying degrees of accuracy. Each component is discussed in turn below.

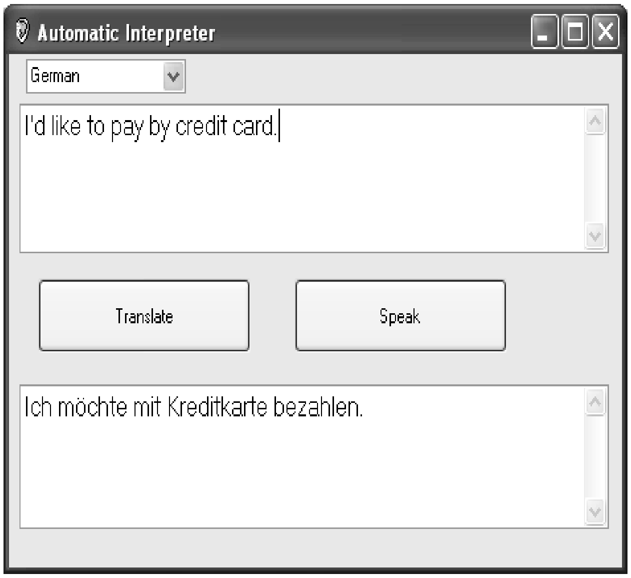
- **Speech Recognition -** With *Dragon Systems Naturally Speaking Standard Edition* (the current version 10 is about $99), a speaker can achieve up to 99% accuracy at 150 words per minute under ideal conditions with no background noises and thorough system training [4]. For example, one study with *Dragon Systems* was able to achieve 97% accuracy with just 30 minutes of user enrollment and training [27]. However, in tests with novice users with little training, poorer results are usually obtained. In a test at PC Labs with IBM's *ViaVoice*, five subjects obtained an accuracy of 89.1% at 74 wpm and 87.8% at 69 wpm, and a study of 12 users speaking 300 words with minimal training found no difference between *DragonNaturally Speaking Version 5.0* and *ViaVoice Version 8.0* (both about 80% accuracy) [3].

- **Language Translation -** *Google Translate* (http://translate.google.com/ *)* currently provides free, automatic translations among any of 51 different languages [14]. Prior research has shown that good translations can be obtained, especially when they are among western European languages such as English, German, and Spanish [1]. However, even if not perfect for many language pairs, the resulting translation is often good enough to know "who did what to whom" [7].

- **Speech Synthesis -** Free, downloadable Lernout and Hauspie Text-To-Speech (L&H TTS) 3000 modules provide male and female voices for Dutch, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, and Spanish for use by the *Microsoft Agent* Active-X control in *Microsoft Visual Studio*. Also, by using the Dutch voices for similar languages (Afrikaans, Danish, Norwegian, and Swedish); Italian for Romanian, Russian for Belarusian, Bulgarian, Ukrainian, and Serbian; and Spanish for Filipino and Galician, 11 additional languages can be supported, for a total of 20. Other language modules such as those for Chinese and Polish can be purchased. In addition, the other 30 non-English languages currently available with *Google Translate* could be accommodated through extra programming by converting the text into phonemes and using one of the 9 installed TTS languages, changing the pitch and speed of the voice as necessary. However, the resulting speech is not likely to seem as natural.

Figure 1 shows an example of the prototype system in use. With *DragonSystems* running in the background, the user selects the destination language in the dropdown box, places the cursor in the top text box, and begins to dictate. After groups of a few words are spoken, the speech recognition software automatically enters the text. If words are not recognized, the user has the option of editing or retyping the transcript. Next, the user clicks the "Translate" key for a translation in the selected language, and then clicks the "Speak" key for the computer to synthesize the foreign vocalization.

Figure 1: Sample screen of the automatic interpreter



German

I'd like to pay by credit card.

Ich möchte mit Kreditkarte bezahlen.

EVALUATION

In an attempt to test the accuracy of the system, an adult, male, American (subject #1) and an adult, female, Indian (subject #2) trained *Dragon Systems Naturally Speaking* thoroughly and then read aloud a sample of 38 lines of travel phrases with 721 words in English from the Fodor's Web site found at http://www.fodors.com/language/ (see Table 1). The BLEU score [18] for subject #1's transcript was 95.6 (maximum for the score is 100), while that for subject #2 was 70.5. The F measure that combines calculations of the recall and precision was 97 for subject #1 and 88 for subject #2.

Table 1: Speech recognition results (bold shows errors)

Original	**Subject 1**	**Subject 2**
How much does that cost?	How much does that cost?	How much does that cost?
At what time does the store open?	At what time does the store open?	**And what time that the store opened?**
At what time does the store close?	At what time does the store close?	**And what time does this do does?**
What would you like?	What would you like?	What would you like?
Can I help you?	**And I help you?**	**Can I had you?**
I would like this.	I would like this.	**I would like this.**
Here it is. Is that all?	Here it is. Is that all?	**year it is 88 is that all?**
I would like to pay in cash.	I would like to pay in cash.	**I would like to be in cash.**
I would like to pay by credit card.	I would like to pay by credit card.	I would like to pay by credit card.
Can I order this online?	Can I order this online?	**Can I ordered this online?**
Where is there a good restaurant?	Where is there a good restaurant?	**Then is there a good restaurant?**
A table for 2, please. The menu, please.	A table for 2, please. The menu, please.	**A table 4 2, please. The menu among please.**
The wine list, please.	The wine list, please.	**Be it the wine list, please.**
I would like something to drink.	I would like something to drink.	I would like something to drink.
A glass of water, please.	A glass of water, please.	A glass of water, please.
Do you have a vegetarian dish?	Do you have a vegetarian dish?	Do you have a vegetarian dish?
I like my steak well done. I like my steak rare.	I like my steak well done. I like my steak rare.	I like my steak well done. I like my steak rare.
I had a wonderful time.	I had a wonderful time.	I had a wonderful time.
This is my girlfriend.	This is my girlfriend.	This is my girlfriend.
This is my boyfriend.	This is my boyfriend.	This is my boyfriend.
This is my friend.	This is my friend.	This is my friend.
Where do you live?	**Where you live?**	**When do you live?**
I live in New York.	I live in New York.	**I gave in New York.**
It's nice to meet you.	It's nice to meet you.	It's nice to meet you.
What is your name?	What is your name?	What is your name?
How are you?	How are you?	How are you?
Is it near?	Is it near?	Is it near?
Is it far?	**Isn't far?**	Is it far?
Go straight ahead.	Go straight ahead.	Go straight ahead.
Go that way.	**Do that way.**	Go that way.

Turn right.	Turn right.	Turn right.
Turn left.	Turn left.	Turn left.
Take me to this address.	Take me to this address.	Take me to this address.
What is the fare?	What is the fare?	What is the fare?
Stop here.	Stop here.	Stop here.
I would like a map of this city.	I would like a map of this city.	I would like a map of this city.
Where are the taxis?	Where are the taxis?	**There are the taxis?**
Where is the subway?	Where is the subway?	Where is the subway?
Where is the exit?	Where is the exit?	Where is the exit?

This SR text was automatically translated into German and Spanish, and an objective evaluator reviewed the results for comprehension on a scale of 0-100%. The mean German accuracy was 99.2% while the Spanish accuracy was 98.5% for subject #1, while that for subject #2 was 94.19% and 94.08%, respectively. Nearly all of the miscomprehension in the final translations arose from the English speech recognition errors. For example, "Be it the wine list, please." and "There are the taxis?" were mostly understood, but "Can I had you?" and "year it is." were less well understood.

Direct comparisons cannot be made because of different source text samples and input and output languages, but these limited results were much better than the 80% accuracy found using German, English, and Japanese in the Wahlster [23] study and also higher than the 89.9% found using English and Chinese in the Wang & Seneff, [25] study. Thus, the system might already be useful, especially if a human interpreter is not available.

CONCLUSION

Automatic speech recognition systems can be implemented using free and commercial software integrated with any of several programming languages (e.g., Javascript, Java, C, or Visual Basic), and results might be even more accurate with such a system than with proprietary or experimental products reviewed earlier. This paper has described a prototype SR-MT system that provides automatic interpretation of spoken unrestricted English sentences to nine other languages, although only German and Spanish were used for testing purposes. Using two English speakers, results showed an average comprehension of 96.5% for both target languages together. We believe this accuracy is adequate for most informal, tourist-related dialogs when the alternatives are no interpretation at all or much slower use of portable electronic devices with limited preprogrammed travel phrases.

The study suffers from several limitations, however. First, only two subjects were used for testing the prototype. A larger sample is necessary for adequate statistical analysis. Second, a restricted set of sample phrases was used. Longer, more complex sentences could result in worse performance. Third, only two languages were evaluated. *Google Translate's* accuracy varies considerably based upon the language-pair combination selected [2]. Fourth, the prototype should be tested in a real-world environment with distracting background noises that are likely to degrade its accuracy.

Future research will investigate the system's accuracy using other language pairs. In addition, with only small changes, we believe completely automated speech-to-speech interpretation can be provided through mobile devices in many languages, unlimited by topic area, with off-the-shelf software.

REFERENCES

1. Aiken, M. and Ghosh, K. (2009). Automatic translation in multilingual business meetings. *Industrial Management & Data Systems*, 109(7), 916-925.
2. Aiken, M., Park, M., Simmons, L., and Lindblom, T. (2009). Automatic translation in multilingual electronic meetings. *Translation Journal*, 13(9), July.
3. Broughton, M. (2002). Measuring the accuracy of commercial automated speech recognition systems during conversational speech. Virtual Conversational Characters (VCC) Workshop, Human Factors Conference, Melbourne, Australia, December.
4. Dragon (2009). Dragon Naturally Speaking. Retrieved January 11, 2010, from http://www.scansoft.com/naturallyspeaking/
5. Frederking, R., Rudnicky, A., Hogan, C., Lenzo, K. (2000). Interactive speech translation in the Diplomat Project. *Machine Translation*, 15(1-2), 27-42.
6. Fujii, K., Yoshino, T., Shigenobu, T., & Munemori, J. (2005). Development of an intercultural collaboration system with semantic information share function, *In*: *Knowledge-Based Intelligent Information and Engineering Systems*, (425-430), Springer: Berlin.
7. Garreau, J. (2009). Tongue in check. Washington Post (DC) (05/24/09).

8. House, J. (2003). English as a lingua franca: A threat to multilingualism? *Journal of Sociolinguistics*, 7(4), 556–578.
9. Jenkins, J. (2007). English as a Lingua Franca: Attitude and identity, Oxford, UK, Oxford University Press.
10. Lim, J. & Yang, Y. (2008). Exploring computer-based multilingual negotiation support for English-Chinese dyads: can we negotiate in our native languages? *Behavior & Information Technology*, 27(2), 139-151.
11. Matsubara, S., Toyama, K., and Inagaki, Y. (1999). Sync/Trans: Simultaneous machine interpretation between English and Japanese. Advanced Topics in Artificial Intelligence. Springer: Berlin.
12. Morimoto, T., Takezawa, T., Yato, F., Sagayama, S., Tashiro, T., Nagata, M., and Kurematsu, A. (1993). ATR's speech translation system: ASURA. *EUROSPEECH 1993 Third European Conference on Speech Communication and Technology*. Berlin, Germany, September 22-25.
13. Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Jin-Song Z., Yamamoto, H., Sumita, E., and Yamamoto, S. (2006). The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2), 365-376.
14. Och, F. (2009). 51 Languages in Google Translate. Google Research Blog. Retrieved November 8, 2009 from http://googleresearch.blogspot.com/2009/08/51-languages-in-google-translate.html
15. Ogden, B. (2005). Computer mediated multilingual translation. Computing Research Lab, New Mexico State University. Retrieved November 8, 2009 from http://crl.nmsu.edu/Research/Projects/ogden/TrIM/
16. Ogura, K., Hayashi, Y., Nomura, S., & Ishida, T. (2004). User adaptation in MT-mediated communication, *Paper presented at the First International Joint Conference on Natural Language Processing (IJCNLP-04),* 596-601.
17. Okumura, A., Iso, K., Doi, S., Yamabana, K., Hanazawa, K., and Watanabe, T. (2002). An automatic speech translation system for travel conversation. *NEC Research and Development*, 43(1), 37-40.
18. Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July, 2002, 311-318.
19. Pastor, M., Sanchis, A., Casacuberta, F., and Vidal, E. (2001). EUTRANS: A speech-to-speech translator prototype. *Proceedings of EuroSpeech.*
20. Peterson, R. (2006). IBM strives for superhuman speech. *PC Magazine,* January 24.
21. Rayner, M. and Bouillon, P. (1995). Hybrid transfer in an English-French spoken language translator. In *4th Congress of the Italian Association for Artificial Intelligence*, Florence, Italy, October 11-13, 153-162.
22. SRI (2009). Speech translation research at SRI international. Full spontaneous translation. Retrieved January 11, 2010, from http://www.speech.sri.com/projects/translation/full.shtml
23. Wahlster, W. (2009). Mobile speech-to-speech translation of spontaneous dialogs: An overview of the final Verbmobil system. Retrieved January 11, 2010, from http://verbmobil.dfki.de/ww.html
24. Waibel, A. (1996). Interactive translation of conversational speech. *Computer*, 29(7), 41-48.
25. Wang, C. and Seneff, S. (2006). High-Quality Speech Translation in the Flight Domain. *Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, September 17-21, 2006
26. Watanabe, T., Okumura, A., Sakai, S., Yamabana, K., Doi, S., and Hanazawa, K. (2000). An automatic interpretation system for travel conversation. *Sixth International Conference on Spoken Language Processing (ICSLP 2000),* Beijing, China, October 16-20.
27. Zafar, A., Overhage, J., and McDonald, C. (1999). Continuous speech recognition for clinicians. *Journal of the American Medical Informatics Association*, 6, 195-204.
28. Zheng, J. (2006). Embedded multilingual mobile applications. *Multilingual,* April/May
29. Zhou, B., D´echelotte, D., and Gao, Y. (2004). Two-way speech-to-speech translation on handheld devices. *Proceedings of the International Conference of Spoken Language Processing (ICSLP*), Korea, Oct.