

Speech to Speech translation for Nurse Patient Interaction

Farzad Ehsani, Jim Kimzey, Elaine Zuber, Demitrios Master
Fluential Inc./ 1153 Bordeaux Dr.,
Sunnyvale, CA 94089
{farzad, jkimzey, elaine, dlm}
@fluentialinc.com

Karen Sudre
TeleNav, Inc./1130 Kifer Road,
Sunnyvale CA, 94086
karens@telenav.com

Abstract

S-MINDS is a speech translation system, which allows an English speaker to communicate with a limited English proficiency speaker easily within a question-and-answer, interview-style format. It can handle dialogs in specific settings such as nurse-patient interaction, or medical triage. We have built and tested an English-Spanish system for enabling nurse-patient interaction in a number of domains in Kaiser Permanente achieving a total translation accuracy of 92.8% (for both English and Spanish). We will give an overview of the system as well as the quantitative and qualitatively system performance.

1 Introduction

There has been a lot of work in the area of speech to speech translation by CMU, IBM, SRI, University of Geneva and others. In a health care setting, this technology has the potential to give nurses and other clinical staff immediate access to consistent, easy-to-use, and accurate medical interpretation for routine patient encounters. This could greatly improve safety and quality of care for patients who speak a different language from that of the healthcare provider.

This paper describes the building and testing of a speech translation system, S-MINDS (Speaking Multilingual Interactive Natural Dialog System), built in less than 3 months with Kaiser Permanente Hospital in San Francisco, CA. The system was able to gain fairly robust results for the domains that it was designed for, and we believe

that it does demonstrate that building and deploying a successful speech translation system is becoming possible and even commercially viable.

2 Background

The number of people in the U.S. who speak a language other than English is large and growing, and Spanish is the most commonly spoken language next to English. According to the 2000 census, 18% of the U.S. population over age 5 (47 million people) did not speak English at home—a 48% increase from 1990. In 2000, 8% of the population (21 million) was LEP (Limited English Proficiency), with more than 65% of that population (almost 14 million people) speaking Spanish.

A body of research shows that language barriers impede access to care, compromise quality, and increase the risk of adverse outcomes. Although trained medical interpreters and bilingual healthcare providers are effective in overcoming such language barriers, the use of semi-fluent healthcare professionals and *ad hoc* interpreters (such as family members and friends) cause more interpreter errors and lower quality of care (Flores 2005).

When friends and family interpret, they are prone to omit, add, and substitute information. Often they inject their own opinions and observations, or impose their own values and judgments, rather than interpreting what the patient actually said. Frequently these *ad hoc* interpreters have limited English capabilities themselves and are unfamiliar with medical terminology. Furthermore, many patients are reluctant to disclose private or sensitive information in front of a family member, thus giving the doctor an incomplete picture; this sometimes prevents a

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (<http://creativecommons.org/licenses/by-nc-sa/3.0/>). Some rights reserved.

correct diagnosis. For example, a battered woman is unlikely to reveal the true cause of her injuries if her husband is being used as the interpreter.

The California Academy of Family Physicians Foundation conducted practice visits in 2006 and found that, “Although they realize the use of family members or friends as interpreters is probably not the best means of interpretation, all practice teams use them.” (Chen et al 2007)

3 System Description

Fluential’s speech translation system, S-MINDS¹, has a hybrid architecture (Figure 1) that combines multiple ASR engines and multiple translation engines. This approach only slightly increases the development cost of new translation applications, but it greatly improves the accuracy and the coverage of the system by leveraging the strengths of both statistical and grammar/rules-based systems. Furthermore, this hybrid approach enables rapid integration of new speech recognition and translation engines as they become available.

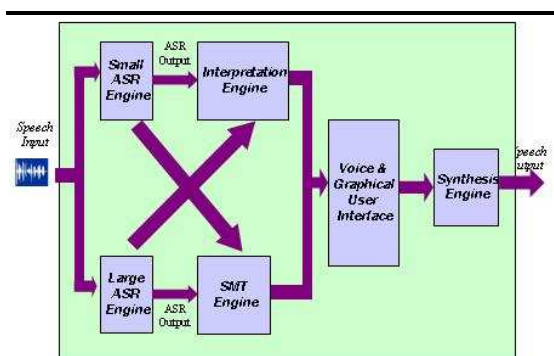


Figure 1. The hybrid system architecture of S-MINDS combines multiple ASR engines with an interpretation engine and an SMT engine. Note that this figure describes the interaction in English to-second language direction only. The 2nd language-to-English direction has only the small ASR engine and the interpretation engine.

3.1 Components of Speech Translation System

S-MINDS has a modular architecture with the components described below. All of these components already exist, so they will not need to be

developed to conduct the research proposed in Phase I.

3.1.1 ASR Engine

S-MINDS employs multiple acoustic engines so the best engine can be chosen for each language. Within each language, two separate language models are active at the same time, telling the ASR engines which words and phrases to recognize. A smaller, more directed language model with higher accuracy is used to capture important and frequently used concepts. For less frequently used concepts, a larger language model that generally has broader coverage but somewhat lower accuracy is used. The combination of these two provides high accuracy for responses that can be anticipated and slightly lower accuracy but broader coverage for everything else. This method also allows development of new domains with very little data—for each domain, only a new domain-specific small language model needs to be built.

3.1.2 Interpretation Engine

Fluential has created an interpretation engine that is an alternative to an SMT engine. The S-MINDS interpretation engine uses information extracted from the output of the ASR engine and then performs a paraphrase translation in semantic space. This process is similar to what human interpreters do when they convey the essential meaning without providing a literal translation.

The advantage of an interpretation engine is that new domains can be added more quickly and with less data than is possible with an SMT engine. For high-volume, routine interactions, an interpretation engine can be extremely fast and highly accurate; however, the translation may lose some of the nuance. Again, this means that highly accurate target applications can be built with very little data—only a few examples of each concept are needed to train the interpretation engine.

3.1.3 Statistical Machine Translation Engine

For the S-MINDS SMT engine, Fluential is developing a novel approach that has generally improved the accuracy of speech translation systems.² This approach capitalizes on the intuition that language is broadly divided into two levels:

¹ Speaking Multilingual Interactive Natural Dialog System

² This effort is ongoing; it has not yet been fully implemented.

structure and vocabulary. Traditional statistical approaches force the system to learn both types of information simultaneously. However, if the acquisition of structural information is kept separate from the acquisition of vocabulary, the resulting system should learn both levels more efficiently. And by modifying the existing corpus to separate structure and vocabulary, we have been able to take full advantage of all the information in the bilingual corpus, producing higher-quality MT without requiring large bodies of training data. The most recent modification to this approach was the use of distance-based ordering (Zens and Ney, 2003) and lexicalized ordering (Tillmann and Zhang, 2005) to allow for multiple language models, including non-word models such as part-of-speech improved search algorithm, in order to improve its speed and efficiency.

3.1.4 VUI+GUI System

S-MINDS has a flexible user interface that can be configured to use VUI only or VUI+GUI for either the English speaker or the second-language speaker. Also, the English speaker can experience a different user interface than the second-language speaker. The system has the flexibility to use multiple types of microphones, including open microphones, headsets, and telephone headsets. Speech recognition can be confirmed by VUI, GUI, or both, and it can be configured to verify all utterances, no utterances, or just utterances that fall below a certain confidence level.

3.1.5 Synthesis Engine

S-MINDS can use text-to-speech (TTS) synthesis throughout the system; alternatively, it can use TTS in its SMT-based system and chunk-based recordings that are spliced together in its interpretation engine. Fluentia licenses its TTS technology from Cepstral, and other vendors. In general we do not expect to be doing any research and development activities in this area, as Cepstral can easily create good synthesis models from the 10 hours of provided speech data (Schultz and Black, 2006, Peterson, 2007).

4 System Building

Fluentia conducted five activities in order to build the system. They included: (1) Defining the task, (2) Collecting speech data to model nurse-patient interactions, (3) Building and testing a speech translation system in English and

Spanish, (4) Using the system with patients and nurses and collecting data to measure system performance, and (5) Analyzing the results.

To define the task, Fluentia conducted a two-hour focus group with six registered nurses from Med/Surg unit of Kaiser Medical Center in San Francisco. In this focus group, the nurses identified six nurse-patient encounter types that they wanted to use for the evaluation. These were: (1) Greeting/Goodbye, (2) Vital Signs, (3) Pain Assessment, (4) Respiratory Assessment, (5) Blood Sugar, (6) Placement of an I.V.

Fluentia then collected speech data over a four-week period by recording nurse-patient interactions involving 11 nurses and 25 patients. Fluentia also recruited 59 native Spanish speakers who provided speech data using an automated system that walked them through hypothetical scenarios and elicited their responses in Spanish.

The English recognizer had a vocabulary of 2,003 and it was trained with 9,683 utterances. The Spanish recognizer had a vocabulary of 822, and it was trained with 1,556 utterances. We suspect that the vocabulary size in Spanish would have been much bigger if we had more data.

5 System Evaluation

After building and testing the speech translation system, Fluentia conducted a two-hour training session for each of the nurses before using the system with patients. A bilingual research assistant explained the study to patients, obtained their consent, and trained them for less than five minutes on the system. Nurses then used the system with Spanish-speaking patients for the six nurse-patient encounters that were built into the system. The system was used by three nurses with eleven patients for a total of 95 nurse-patient encounters creating a total of 500 conversation segments.³

To protect patients from a mistranslation, each encounter was remotely monitored by a bilingual interpreter, who immediately notified the nurse any time the system mistranslated. Each encounter was recorded, transcribed, and translated by a human.

3.1 Scoring Accuracy

³ A conversation segment is a single continuous sequence of speech in a single language plus the translation of what was said.

The human translations were compared to the system's translations and given a score using the Laws Methodology of either Good, Fair, Poor, Mistranslated, or Not Translated. (Laws, 2004). If a translation were scored as Good or Fair, it was considered accurate. If the translation were scored as Poor, Mistranslated, or Not Translated, it was considered inaccurate.

Table 2 and 3 give examples of how we have used Law's method to grade actual interaction results from nurses and patients.

What Nurse Said	S-MINDS™ Translation	Human Translation	S-MINDS Accuracy
I will give you an I.V.	Voy a colocarle un cateter para liquidos intravenosos.	Voy a colocarle un cateter de liquidos intravenosos.	Good
Let me check if I can give you medication for that.	Dejeme chequear si puedo darle algun medicamento.	Permitame reviso si puedo darle algun medicamento para eso.	Fair
I will check your...	Yo voy a revisarle los vendajes	Voy a revisar su ...	Poor
Did someone take your vitals?	¿Le tomare sus signos vitals?	¿Alguien tomo sus signos vitals?	Mistranslated
Your heart rate is normal.	---	Su frecuencia cardiaca es normal.	Not Translated

What Patient Said	S-MINDS Translation	Human Translation	S-MINDS Accuracy
No, no tengo dificultad en respirar.	I don't have difficulty breathing.	No, I don't have difficulty breathing.	Good
En la parte baja del estomago.	The lower part of my stomach.	In the lower part of my stomach.	Fair
N/A	N/A	N/A	Poor
N/A	N/A	N/A	Mistranslated
Los huesos.	---	My bones.	Not Translated

6 Results

- Our internal milestones for Phase I was to achieve 80% accuracy using the Laws Methodology. Out of 500 conversation segments, the

speech translation system had an overall accuracy rate of 93% combining both nurse- and patient-conversation segments,

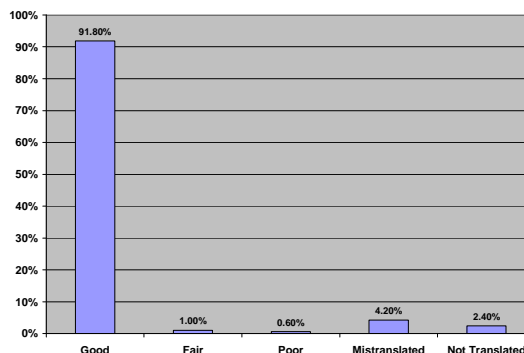


Figure 2: Total results for both nurses and patients.

6.1 Nurse Translation Results

Looking at just nurse conversation segments, the speech translation system had higher accuracy than for patient segments. Out of 404 nurse segments, the speech translation system had a 94% accuracy rate.

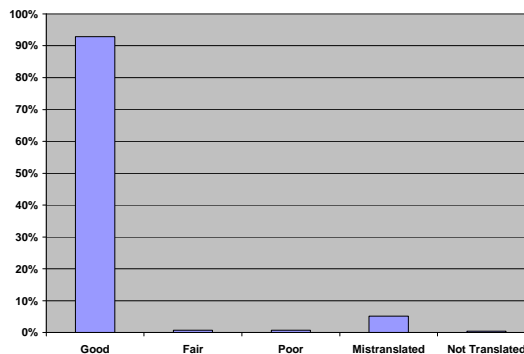


Figure 3: Accuracy for Nurse Conversational Segments

The biggest problem with system performance with nurses was with mistranslations. When nurses tried to say things that were not in the system, the system tried to map their utterances to something that was in the system. In each case of mistranslation, the system told the nurse what it was about to translate, gave the nurse a chance to stop the translation, and then translated the wrong thing when the nurse did not respond. We believe that system performance can be greatly improved in by collecting more speech data from patients and nurses, making changes to the user interface, and improving our training program.

6.2 Patient Translation Results

Looking at just patient conversation segments, the speech translation system had lower overall accuracy than for nurse segments. Out of 96 patient segments, the speech translation system had a 90% accuracy rate.

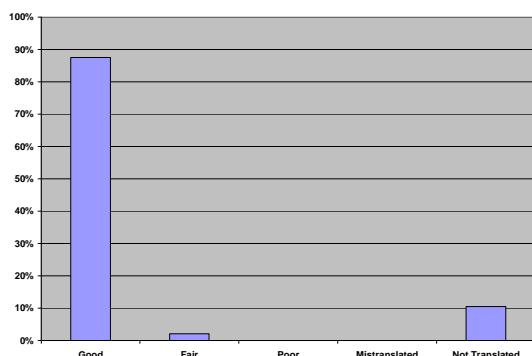


Figure 4: Results for Patients

All of the problems with system performance with patients were with responses that the system was not able to translate. The system never gave a Poor translation or Mistranslated. So there were times when the nurse knew that the patient tried to say something that the system could not translate, but there was never a time when the system gave the nurse false information. However, this percentage is quite high, and in a large context, it might cause additional problems.

6.3 Nurse Survey Results

After each time using the system, the nurses completed a user satisfaction survey that had five statements and asked them assign a 1-to-5 Likert score to each statement with 1 meaning “Strongly Disagree” and 5 meaning “Strongly Agree.” Average scores for each question were:

4.7 The speech translator was easy to use.

4.5 The English voice was fluent and easy to understand.

4.4 I understood the patient better because of the speech translator.

4.5 I feel that I am providing better medical care because of the speech translator.

4.7 I would like to use the speech translator with my patients in the future.

6.4 Patient Survey Results

The patients also completed a similar user satisfaction survey, translated to Spanish, after using

the system. Their average scores for each question were:

4.6 The speech translator was easy to use.

4.8 The Spanish voice was fluent and easy to understand.

4.7 I understood my nurse better because of the speech translator.

5.0 I feel that I am receiving better medical care because of the speech translator.

4.9 I would like to use the speech translator with my nurse in the future.

6.5 ANOVA Testing

We conducted Analysis of Variance (ANOVA) testing to evaluate whether there were any significant variations in translation accuracy by patient, nurse, or encounter type. There were no significant differences.

7 Discussion

We were able to build and evaluate a system in 3 months and show its utility by nurses and patients in clinical setting. The system seemed to work and was liked by both nurses and patients. The next question is whether such a system can scale and cover a much larger domain, and how much data and training is required to accomplish this.

References

- Chen A., et al. (2007), *Addressing Language and Culture—A Practice Assessment for Health Care Professionals*, p3.
- Flores Glenn, (2005), “The Impact of Medical Interpreter Services on the Quality of Health Care: A Systematic Review,” *Medical Care Research and Review*, Vol. 62, No. 3, pp. 255-29
- Laws, MB, Rachel Heckscher, Sandra Mayo, Wenjun. Li, Ira Wilson, (2004), “A New Method for Evaluating the Quality of Medical Interpretation,” *Medical Care*. 42(1):71-80, January 2004
- Peterson Kay (2007). Senior Linguist, Cepstral LLC, Personal Communication.
- Schultz, Tanja and A. W Black (2006), “Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs”

Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2006), Toulouse, France, May 15-19, 2006.

Tillmann, Christoph and T. Zhang, (2005), "A localized prediction model for statistical machine translation," in *Proceedings of the 43rd Annual Meeting of the ACL*, pp. 557-564, Ann Arbor, June 2005.

Zens, Richard, and H. Ney, (2003), "A comparative study of reordering constraints in statistical machine translation," in *Proceedings of the 41st Annual Meetings of the ACL*, pp. 144-151, Sapporo, Japan, July 2003 Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.