Journal of Diabetes Science and Technology
Volume 6, Issue 4, July 2012

ORIGINAL ARTICLE

Performance of a New Speech Translation Device in Translating Verbal Recommendations of Medication Action Plans for Patients with Diabetes

R. William Soller, Ph.D., Philip Chan, Pharm.D., and Amy Higa

Abstract

Background:

Language barriers are significant hurdles for chronic disease patients in achieving self-management goals of therapy, particularly in settings where practitioners have limited nonprimary language skills, and in-person translators may not always be available. S-MINDS© (Speaking Multilingual Interactive Natural Dialog System), a concept-based speech translation approach developed by Fluential Inc., can be applied to bridge the technologic gaps that limit the complexity and length of utterances that can be recognized and translated by devices and has the potential to broaden access to translation services in the clinical settings.

Methods:

The prototype translation system was evaluated prospectively for accuracy and patient satisfaction in underserved Spanish-speaking patients with diabetes and limited English proficiency and was compared with other commercial systems for robustness against degradation of translation due to ambient noise and speech patterns.

Results:

Accuracy related to translating the English–Spanish–English communication string from practitioner to device to patient to device to practitioner was high (97–100%). Patient satisfaction was high (means of 4.7–4.9 over four domains on a 5-point Likert scale). The device outperformed three other commercial speech translation systems in terms of accuracy during fast speech utterances, under quiet and noisy fluent speech conditions, and when challenged with various speech disfluencies (i.e., fillers, false starts, stutters, repairs, and long pauses).

Conclusions:

A concept-based English–Spanish speech translation system has been successfully developed in prototype form that can accept long utterances (up to 20 words) with limited to no degradation in accuracy. The functionality of the system is superior to leading commercial speech translation systems.

J Diabetes Sci Technol 2012;6(4):927-937

Author Affiliation: Center for Self-Care, School of Pharmacy, University of California, San Francisco, San Francisco, California

Abbreviations: (CI) confidence interval, (EMR) electronic medical record, (LEP) limited English proficiency, (MAP) medication action plan, (MTM) medication therapy management, (UCSF) University of California, San Francisco

Keywords: accuracy, diabetes, electronic speech translation, medication therapy management, speech translation system

Corresponding Author: R. William Soller, Ph.D., University of California, San Francisco, School of Pharmacy, 521 Parnassus Ave., Box 0622, San Francisco, CA 94143-0622; email address sollerrw@pharmacy.ucsf.edu

Introduction

Patients with limited English proficiency (LEP) are a challenging segment of the U.S. population.[1] Limited English proficiency patients report decreased satisfaction in communicating with health care providers and may be less likely to understand medical situations, be scheduled for follow-up appointments, or receive informed consent.[2–7] Nationwide quality assessments have consistently noted less favorable reports of care among LEP patients.[8–11]

Limited English proficiency is an independent predictor of poor glycemic control among insured U.S. Latinos with diabetes, an association not observed when care is provided by language-concordant physicians.[12] This is significant, as diabetes affects over 23 million Americans, and Hispanics have a disproportionately higher prevalence of diabetes (10.4%) and higher rates of diabetes-induced end-stage renal disease and mortality from diabetes as non-Hispanic whites.

In over 8000 U.S. locations, community health clinics serve 23 million patients and provide one-quarter of all primary care visits for the nation's low-income population.[13–15] Nationally, over 30% of community clinic patients have LEP, with 20% of clinics having more than 50% LEP patients.[16] Approximately 40% of clinics report "translation assistance" as "very important" to patient care, with 30% of visits requiring an extra 16–30 min.

California health workers are predominantly Caucasian and do not reflect the ethnic diversity of the state's population. Latinos are significantly underrepresented throughout the health workforce, with only 4% of physicians and 4% of registered nurses being Latino, while over 30% of the state population is Latino.[17–19] Twenty-nine percent of community health centers pay bilingual staff additional compensation specifically to provide interpretation services as well as other job duties.

On this background, Fluential Inc. and the University of California, San Francisco (UCSF) School of Pharmacy Center for Self-Care collaborated to demonstrate the usefulness of Fluential's English-to-Spanish speech translation system (S-MINDS©—Speaking Multilingual Interactive Natural Dialog System) for Spanish-speaking LEP patients with diabetes. S-MINDS is a concept-based speech translation approach that can be applied to bridge technologic gaps limiting the complexity and length of utterances that can be recognized and translated by devices. Since 2006, the center has provided medication therapy management (MTM) services to underserved patients with diabetes at St. Anthony Medical Clinic. Comprehensive literature reviews have shown pharmacist-mediated MTM services provide clinically significant improvements in clinical, economic, and humanistic outcomes in a variety of disease states and settings.[20–22]

The research objectives were to (1) assess the accuracy of S-MINDS in patient visits, (2) determine patient satisfaction, and (3) compare the functionality of S-MINDS to current commercial speech translation systems for accuracy of translation in different audio environments and ability to overcome speech disfluencies (e.g., fillers, false starts, stutters) and handle rapid utterances.

Methods

This was a prospective assessment of each communication step of the prototype system with documentation of accuracy, patient satisfaction, and laboratory comparisons with other translation systems.

Prototype Development

For testing the feasibility of adapting S-MINDS to clinical settings, the system was deconstructed to test each of its communication steps for accuracy. In practice, S-MINDS users would have the option to omit or view the text-recognition portions for confirming translations. The communication steps of S-MINDS are (1) speech initiation of a medication-related recommendation by the English-speaking practitioner, (2) visual choice by the English-speaking practitioner of the correct corresponding English sentence from three options on the smart phone screen, (3) audio Spanish-language translation of the visual choice by S-MINDS, (4) verbal restatement of the audio sentence by the Spanish-speaking LEP patient, (5) choice by the patient from three Spanish-language written options on the smart phone screen, and (6) audio English translation by S-MINDS which, if accurate translation occurred, would be the medication-related recommendation stated in step 1.

Content, speech recognition, and translation were built into S-MINDS, with a focus on medication action plans (MAPs), which summarize each patient visit into several tangible steps for patient application in their own self-care. Often, those steps are written as complete sentences for the patient on a hard copy page. Twenty-one single-concept MAP recommendations were selected from 25 LEP patient visits with two MTM-trained pharmacists and over 1000 visits from our other clinical services. These recommendations included topics of glycemic control ($n = 4$), medication administration ($n = 3$), adherence ($n = 2$), medication change ($n = 1$), education ($n = 5$), side effects ($n = 1$), laboratory tests ($n = 2$), lifestyle ($n = 2$), and physician referrals ($n = 1$). Also, 12 MAP recommendations paired a stem phrase relating to glycemic control ("test your blood sugar") with a modifier (e.g., "at least once a week," "more often when you are sick," "before dinner and at bedtime," or "once when you wake up and 1 to 2 h after each of your three meals" among others, all written at a fifth-grade reading level (**Appendix A**).

Assessment of Accuracy

Patients were included if they were current patients in the UCSF–St. Anthony Diabetes Telepharmacy Clinic, were 21–85 years of age, had a physician diagnosis of type 2 diabetes, were prescribed one or more oral or injectable diabetes medicines, had electronic medical record (EMR) documentation of LEP status and patient preference of Spanish for medical visits, and had completed a consent form. The testing of the S-MINDS system was in a private office of the clinic. Patients were given a $20 incentive.

Accuracy was documented by two researchers who were fluent in medical English and Spanish and the practitioner. Verbal and written statements at each communication step were categorized as (a) correct, if the statement was an exact representation of the stated or written statement; (b) conceptually correct, if the statement had the same meaning as determined by the researcher (e.g., I will test [versus check] my blood sugars twice a day); (c) partial, if only a portion of the statement was represented correctly; or (d) incorrect, if the entire statement was a misrepresentation. Researchers recorded if patients indicated they forgot what was verbalized by the smart phone or if accidental touchpad errors were made.

Accuracy scores of verbal and audio communication steps (i.e., 1, 3, 4, and 6) were analyzed as raw and corrected scores. If cognitive difficulties (e.g., forgetfulness) resulted in a patient not able to restate an accurate English-recognition-to-Spanish-translation MTM recommendation, then S-MINDS was not documented as being in error and the score was adjusted as correct.

Patient Satisfaction

Patients enrolled in the accuracy assessment (discussed earlier) were asked to complete an anonymous satisfaction survey written in Spanish and formatted as a five-point Likert scale (i.e., totally agree, somewhat agree, uncertain, somewhat disagree, or totally disagree with a series of statements [discussed later]). Statistical analysis involved standard descriptive methods (mean, standard deviation) and derivation of 95% confidence intervals (CIs) of proportions. The study was conducted under an approved institutional review board application of the UCSF Committee on Human Research.

Comparative Laboratory Assessment

Four automatic speech recognition and three machine translation systems were compared independently and with Fluential's concept-based translation processing performance integrated with other systems' automatic speech recognition component as input (see **Table 1** for system configurations). All applications run under Apple's iOS operating system. All devices were pretested to accommodate adaptation to the speaker by the automatic speech recognition systems.

Table 1.
System Configuration for the Tested Speech Translation Devices

Application	Automatic speech recognition domain	Machine translation domain	Processing location
Fluential[a] UCSF iPhone App. v0.1	Pharmacy	Pharmacy	Server
Dragon Dictation iPhone App. v2.0.11	General	—	Server
Jibbigo iPhone App. v1.13171	Travel and medical	Travel and medical	Phone
Google Translate iPhone App. v1.1.1.1731	General	General	Server
Dragon automatic speech recognition + Fluential concept and translate	General	Pharmacy	Server
Jibbigo automatic speech recognition + Fluential concept and translate	Travel and medical	Pharmacy	Phone and server
Google automatic speech recognition + Fluential concept and translate	General	Pharmacy	Server

[a] Fluential S-MINDS speech translation system.

Test conditions using the same male native English speaker were (1) quiet environment, fluent speech (quiet–fluent); (2) noisy environment, fluent speech (noisy–fluent); (3) quiet environment, disfluent speech (quiet–disfluent). Systems were tested at the same time to control for changes in the input. The devices were in a fixed position in relation to the speaker during all tests to control for the effects of microphone position. For the quiet–fluent condition, 102 utterances between 2 and 25 words in length (total of 708 words) were spoken at a rate of 135 words per minute. For the noisy–fluent condition, background noise was added to the quiet–fluent in the form of human speech from a single audio book. Human speech is a particularly challenging type of background noise for automatic speech recognition.

Noise level measurements were taken for quiet room (-54 dB), noise source (-63 dB), and tester's voice (-72 dB). Decibel levels were measured using the iPhone application Decibels. A subset of 34 of the 102 utterances spoken at 157 words per minute (total of 388 words) from the quiet–fluent condition was used. Short utterances such as "at night" were excluded because the effects of noise are not evident on short utterances.

For testing disfluent speech in the quiet environment, 35 utterances of the 102 from the quiet–fluent condition (490 total words) were spoken at 153 words per minute for the disfluencies and at 263 words per minute for the rapid speech. The different types of disfluencies are shown in **Table 2**. Although rapid speech is not technically regarded as disfluency, it was included because it poses similar challenges for automatic speech recognition and requires similar testing. As in the noisy–fluent condition, short utterances were excluded. Speakers are rarely disfluent on short one- or two-word utterances, as these utterances are not challenging to produce and they do not provide enough locations for disfluencies to occur.

Table 2.
Types of Disfluencies Inserted Under Quiet–Fluent Conditions

Type of disfluency	Count	Example/explanation
Fillers	13	Test, uh, your ..., Needs to be, um ...
False starts	5	When—when you plan your ...
Stutters	14	T-t-test your ...
Repairs	21	Next month—I mean—next week ...
Long pause	1	2–3 s pause
Rapid speech	5	Spoken very quickly but clearly

Translation quality assessment was based on the method described by Laws and colleagues.[23] A Spanish–English bilingual linguist coded each conversational segment. A conversational segment was defined as the language spoken during a button press on the device. Each segment was assigned a nominal code shown in **Table 3** that corresponds to codes 1–13 specified by Laws and colleagues.[23] Nominal scores were converted to an ordinal quality score based on a five-point scale with good = 1, fair = 2, poor = 3, mistranslation = 4, or not translated = 5. Word error rate was calculated per Jurafsky and Martin.[24]

Results

Assessment of Accuracy in Underserved Patients with Diabetes

Twenty-one patients met inclusion criteria. They were mainly women (60%) on multiple medications with

Table 3.
Translation Codes

Nominal code[a]	Ordinal quality score
1.0 Literal or fully preserves essential meaning	Good
2.0 Attempted literal, inconsequential syntax error, etc.	Fair
3.0 Paraphrase fully preserves meaning	Good
4.0 Edited report contains literal content	Good
5.0 Report or paraphrase with minor omission or substitution	Fair
6.0 Attempted literal with consequential language error	Poor
7.0 Edited report with significant omission	Poor
8.0 Edited report with addition	Fair
8.1 Edited report with clarifying addition	Good
8.2 Edited report with addition changing meaning	Poor
9.0 Edited report with substitution	Poor
10.0 Edited report with multiple omit, substitution, and/or addition	Poor
11.0 Essentially false[b] report or fabrication	Mistranslation[b]
12.0 No translation[c]	Not translated[c]
13.0 Other or unclassifiable	—

[a] Codes 1–13 from Reference 22.
[b] Changed from "false" in original to "mistranslation."
[c] Changed from "none" in original to "not translated."

virtually no English proficiency in speech or written language (**Table 4**). A total of 33 unique MAP recommendations were tested using the teach-back method (mean 29, range 12–50). Each patient was permitted three attempts be successful in the teach-back step (i.e., stating the Spanish version of the practitioner recommendation). A total of 379 recommendations were tested over 608 attempts by 21 patients.

Table 4.
Demographics

Patients	*n*
Physician diagnosis of type 2 diabetes[a]	21
Spanish as a primary language[a]	21
LEP[b]	21
Male/female	6/15
Duration of type 2 diabetes	3–21 years
Medications	
Single oral medication	2
Multiple oral medication	7
Insulin alone	3
Oral medication(s) plus insulin	9
Mean number of chronic conditions (standard deviation)	5 (1.6)
MAP recommendations	*n*
Number of unique MAP recommendations tested	33
Average number of recommendations tested/patient (standard deviation)	29 (10)
Median number of recommendations tested/patient (range)	27 (12-50)
Total number of recommendations tested	379
Total number of attempts per recommendation tested (each patient was allowed up to three tries per recommendation)	608

[a] Diagnosis in the EMR of St. Anthony Medical Clinic.
[b] Includes patients able to understand limited conversational English but who prefer Spanish for all medical and pharmaceutical encounters, as determined in the St. Anthony Medical Clinic–UCSF diabetes clinic and documented in the EMR.

Table 5 shows the categories of recommendations tested. On average, recommendations that were related to glycemic control, lifestyle, and side effects trended to a somewhat higher average number of attempts than other categories; however, the means for average tries were not significantly different.

Table 5.
Categories of Diabetes Medication Therapy Management Recommendations (Reco's)

Reco categories	Reco's tested[a]		Tries	
	n	%	*n*	Average tries/Reco category
Medication therapy				
Glycemic control	135	36%	224	1.7
Administration	52	14%	65	1.3
Side effects	32	8%	61	1.9
Adherence	24	6%	27	1.1
Medication change	24	4%	32	1.3
Medication education	49	13%	81	1.7
Laboratories	42	11%	60	1.4
Lifestyle (diet/exercise)	22	6%	47	2.1
Referral to physician	8	2%	11	1.4
Total *n*	379	100%	608	100%

[a] The number of unique MAP recommendations was 33 for 21 patients; not all recommendations were tested in all patients.

The accuracy of different components of the communication string is shown in **Table 6** for the voice translation components and in **Table 7** for the written recognition components.

Translation accuracy was 100% for the pharmacist speaking English to the device followed by playing an audio Spanish translation of the pharmacist's MAP recommendation (**Table 6**). Raw accuracy scores were 91% and 90%, respectively, for (a) the patient speaking, in Spanish, the Spanish MAP recommendation that he/she heard from the device in Spanish and (b) the device's audio English translation of what the patient said in Spanish. Raw score was computed using scores including all spoken but not written components of the communication string, including scores defined as "patient forgot," meaning no meaningful Spanish language statement was spoken. If the patient forgot what to say after listening to the device in Spanish, the device had no input to translate at this stage of its development. This represented a conservative approach to defining the overall accuracy score for the device. Adjusted score was computed using scores excluding the "patient forgot" responses in the communication string, which represent patients' unsuccessful attempts to repeat the device's Spanish audio statement due to cognitive difficulty.

Table 6.
Practitioner-to-Patient Accuracy of Communicating Recommendations of Medication Action Plans Using Fluential's S-MINDS Speech Translation System[a]

Score	Physician speaks English to device		Device audio of Spanish translation		Patient speaks Spanish they heard and remembered			Device audio of English translation		
	N	%	*n*	%	*n*	%	95% CI	*N*	%	95% CI
Raw scores										
Total correct	608	100%	608	100%	551	91%	0.88–0.92	546	90%	0.87–0.91
Correct	608	100%	608	100%	399	66%	0.61–0.69	396	65%	0.61–0.68
Partial	0	—	0	—	152	25%	0.21–0.28	150	25%	0.21–0.28
Patient forgot	0	—	0	—	49	8%	0.06–0.10	49	8%	0.06–0.10
Incorrect	0	—	0	—	8	1%	<0.01–0.02	13	2%	0.01–0.03
Adjusted scores										
Total correct	608	100%	608	100%	551	99%	0.97–0.99	546	98%	0.96–0.99
Total incorrect	0	—	0	—	8	1%	<0.01–0.03	13	2%	–0.04

[a] Total number of tries for the 379 different MAP recommendations = 608.

Table 7.
Written Translations on Smart Phone Screen from English-Speaking Practitioner or Spanish-Speaking Patient[a] Using the Fluential S-MINDS Speech Translation System

Cohorts	First option listed		Second option listed		Third option listed		Patient forgot		Not assessed		Error	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Written English recommendations as translated from English spoken by pharmacist												
Display, English *n* = 608	603	99%	2	<1%	0	0	0	0	3	<1%	0	0%
Written Spanish recommendations as translated from Spanish spoken by patient												
Display, Spanish *n* = 608	514	85%	21	3%	10	2%	45	7%	6	<1%	12	2%
Exclude "forgot" *n* = 563	514	91%	21	4%	10	2%	—	—	6	1%	12	2%

[a] See text for 95% CI.

The main reason that the raw scores for the patient component of the communication string was lower than the pharmacist component was the patient did not remember everything that was played in Spanish by the device. Either the patient forgot (8% of tries) or gave a partial answer. If the patient spoke part of the phrase, it was coded as partially correct ("partial" in **Table 6**). All such partials were corrected primarily by a second attempt (28%) and, if needed, by a third attempt (11%, not shown in the table). The partials were easily picked up by the pharmacist making the initial English language recommendations, either by patients acknowledging they had forgotten and/or asking for a retry or in the final English audio statement from the device. The total number of absolutely incorrect full communication strings was approximately 2%—or an overall 98% accuracy score. These are correctable by next-generation device programming.

During the communication string, there were two opportunities for the pharmacist and the patient to verify that what they had spoken in their native language (English or Spanish, respectively) was correct, by choosing from a list of up to four options on the screen of the iPhone (**Table 7**). Overall, the correct phrase was listed as the first, second, or third option in over 99% of attempts for the pharmacist component of the communication string (95% CI: 0.98 to 0.9) and in 90% of attempts for the patient component (95% CI: 0.81 to 0.87). If the attempts where the patient forgot the phrase are excluded (n = 45), then the correct options were listed in first, second, or third place in 97% of the attempts (95% CI: 0.95 to 0.98). This latter number that excludes forgotten phrases is likely the appropriate written verification estimate for the device itself, because forgotten phrases are not spoken. The error rate was 0% for the screen-displayed pharmacist component of the communication string and approximately 2% for the screen-displayed patient component.

Patient satisfaction mean scores were ≥4.7 on a 5-point Likert scale (**Table 8**). All except two patients "totally agreed" that the sound in Spanish from the device was easy to understand. One patient rated it 3 (unsure); however, this patient had EMR-documented cognitive problems associated with age and education. Of the 15 recommendations presented to her, she was unable to complete two communication strings because she forgot (or did not process) the information and correctly answered 13 (86%; 2 on the first try). The other patient "somewhat agreed" that the Spanish-language audio of the device was easy to understand. This patient was unable to complete one communication string because he forgot the information and successfully completed 22 (96%). All other patients who "totally agreed" the device was easy to understand achieved a 97% rate of correct communication strings.

Table 8.
Patient Satisfaction with the Fluential S-MINDS Speech Translation System (n = 21)

Please rate the extent to which you agree with the following statements[a]	Mean (standard deviation)
When I receive instructions from my doctor, nurse, or pharmacist, I prefer to have an interpreter to translate from Spanish to English.	4.7 (0.7)
The sound in Spanish from the device was easy to understand.	4.9 (0.5)
The device is easy to use.	4.7 (0.7)
I have a better understanding of the pharmacist recommendations through the device's translations.	4.7 (0.7)
If an interpreter was not available, I would use this device to help me in most of my medical care.	(0.7)

[a] Likert scale: 1 = totally disagree; 2 = somewhat disagree; 3 = unsure; 4 = somewhat agree; 5 = totally agree.

Comparative Technical Assessment of the Prototype Commercially Available Devices

Rapid speech was particularly difficult for all systems except Fluential's. **Table 9** shows the results of rapidly speaking a common diabetes-related recommendation: "Test your blood sugar at least once a week." In perspective, speech speed for audio books is 150 words per minute and, for speech debaters, ≥350 words per minute.[25]

The Fluential S-MINDS system outperformed the other three commercial systems for speech recognition in both the quiet–fluent and the noisy–fluent conditions as well in the disfluency test (**Table 10**).

Table 9.
Verbatim Output of "Test Your Blood Sugar at Least Once a Week" Spoken at 250 Words per Minute

System	Automatic speech recognition output
Fluential[a]	"Test your blood sugar at least once a week a"
Dragon	"Touch of blusher at least once we get"
Jibbigo	"Tested lecture at least once we get"
Google	"Tester butcher least one hundred twenty"

[a] Fluential S-MINDS speech translation system.

Table 10.
Comparison of Automated Speech Recognition Word Error Rates[a] for Common Practitioner Recommendations Relating to Glycemic Control (Spoken Words to Text)

System	Cleansound setting	Noisy sound setting	Disfluent sound setting
Fluential[b]	0.74%	6.40%[c]	13.54%
Dragon	10.25%	25.46%[c]	27.29%
Jibbigo	12.47%	19.95%[c]	30.29%
Google	12.47%	15.37%[c]	29.71%

[a] Automatic speech recognition word error rate is defined as researcher's speech appearing as text on the iPhone screen. See Methods.
[b] Fluential S-MINDS speech translation system.
[c] $p < .0001$ for all differences between Fluential S-MINDS speech translation system and each of the three commercial systems in each sound environment (unpaired t test). Fluential and Dragon were run together and Jibbigo and Google were run together.

The translations of each system were evaluated by a human observer using the scoring method of Laws and colleagues.[23] **Table 11** shows results for the stand-alone systems and those applying Fluential's concept translations to the automatic speech recognition text of the other

Table 11. Translation Accuracy Comparison[a]

System	Quiet (n = 102) mean(p_1) [p_2]	Noisy (n = 34) mean(p_1) [p_2]	Disfluent (n = 35) mean(p_1) [p_2]
Fluential[b]	1.03	1.00	1.24
Jibbigo	2.06 (<0.0001)[c]	3.38 (<0.0001)[c]	3.6 (<0.0001)[c]
Google	2.21 (0.0001)[c]	2.94 (<0.0001)[c]	3.7 (<0.0001)[c]
Dragon[d] automatic speech recognition and Fluential translation	1.34 (0.004)[c]	1.59 (0.003)[c]	1.34 (not significant)[c]
Jibbigo automatic speech recognition and Fluential translation	1.27 (0.006)[c] [0.0001][e]	1.29 (0.07)[a] [0.0001][e]	1.71 (0.03)[b] [0.0001][e]
Google automatic speech recognition and Fluential translation	1.43 (0.0002)[c] [0.0001][e]	1.26 (0.53)[a] [0.0001][e]	1.77 (0.05)[b] [0.0001][e]

[a] Based on the Laws and colleagues[23] score ratings of 1 to 5: 1, good; 2, fair; 3, poor; 4, mistranslated; 5, no translation. A rating of 1.0 (e.g., Fluential) means all translations were correct. P_1 values are calculated in comparison with the Fluential system. P_2 compares "System automatic speech recognition and Fluential translation" to "System automatic speech recognition and System translation."
[b] Fluential S-MINDS speech translation system.
[c] (p_1) p value for significance of difference between each system (alone or combined with Fluential) versus Fluential alone.
[d] Dragon is not a translation system and was only evaluated on automatic speech recognition performance.
[e] [p_2] p value for significance of difference between Jibbigo and Fluential versus Jibbigo alone or Google and Fluential versus Google alone.

systems. Fluential was statistically significantly better than Google and Jibbigo (note Dragon is not a speech-translation system) and, when combined with the other commercial systems, improved them compared with their stand-alone scores. Fluential alone scored highest across all sound environments.

Discussion

S-MINDS demonstrated 98% accuracy based on adjusted scores to correct for patients' forgetfulness. The system is robust in relation to rapid word rate, word error rate in quiet and noisy settings, concept error rate, and speech disfluencies. It outperformed other commercially available speech translation systems. This is a result of the concept-based approach built into S-MINDS and the fact that the system is programmed specifically for practitioner–patient communication in diabetes.

While the prototype has accuracy, some patients needed all or portions of the recommendations to be repeated. About half the patients had one or more episodes of forgetfulness at some point in the communication string. Sentence lengths were not overly long or complex, and MAP recommendations were based on actual LEP counseling sessions. Hence, the need to repeat recommendations may have been a result of the lower educational level of LEP patients, the lower cognitive functioning of two participants, none engaging in a full counseling session that would have given context and prior mention of MAP recommendations, a number of patients hearing the concepts for the first time (given that they had not progressed to the stage of diabetes that would be associated with some of the MAP recommendations), and/or wavering concentration for unknown reasons during the testing session. It is relevant in this regard that studies show 40–80% of medical information provided by health care practitioners is forgotten immediately.[26,27]

Limitations

A limitation of the study relates to its generalizability, given the relatively low number of underserved patients and practitioners from one clinic. However, the study was part of the first research phase to show the feasibility of further developing the prototype. Further, the patients represented the community clinic LEP patients with diabetes who are the target population for S-MINDS.

There is the potential for rater bias in feasibility studies. However, the raters are University based, are not employees of Fluential Inc., have no economic link to the company except as a subcontract on the National Institutes of Health grant, and did not create the translation system. Further, three raters evaluated in real time each utterance exchange during the simulated clinic sessions. There was disagreement in less than 2% of tries

(n = 12) by patients during the counseling exchanges, with resolution by majority vote. In such instances, the specific text statements and verbal utterances, all of which were recorded, were reviewed to make the final decision regarding accuracy.

If S-MINDS is used in practice as tested (i.e., using each component of the translation system to validate accuracy), then it might seem cumbersome in day-to-day practice. However, the vision is that the majority of the counseling with S-MINDS would be oral, without the written text. The text components would, nevertheless, be available on the system if there was a need to check accuracy.

The MTM recommendations are summarized in the critical teach-back step. It helps ensure patients can at least verbalize the action steps of therapy. The MTM teach-back usually is a concept exchange based on carefully worded sentences from the written MAP. While we selected only the most common recommendations for feasibility testing, they were based on over 1000 LEP counseling sessions, and the high accuracy in this simulated setting supports further work to develop the prototype into a commercial system.

Conclusions

A prototype English–Spanish speech translation system for MTM counseling in diabetes has been successfully developed in a feasibility study of a small sample of underserved LEP patients. It accepts long utterances with limited to no degradation in accuracy, has high patient satisfaction, and performs well in simulated clinic-based scenarios and laboratory comparisons with other commercial systems.

Funding:

This work was funded by the National Institutes of Health (DK089900-02).

Disclosures:

R. William Soller and Philip Chan were supported by a National Institutes of Health sub-award from Fluential Inc.

Acknowledgments:

The authors recognize Farzad Ehsani, Demitrios Master, and the rest of the technical team of Fluential Inc. for their work in the technical assembly of the Fluential speech translation device.

References:

1. Schenker Y, Lo B, Ettinger KM, Fernandez A. Navigating language barriers under difficult circumstances. Ann Intern Med. 2008;149(4):264–9.
2. Morales LS, Cunningham WE, Brown JA, Liu H, Hays RD. Are Latinos less satisfied with communication by health care providers? J Gen Intern Med. 1999;14(7):409–17.
3. Carrasquillo O, Orav EJ, Brennan TA, Burstin HR. Impact of language barriers on patient satisfaction in an emergency department. J Gen Intern Med. 1999;14(2):82–7.
4. Wilson E, Chen AH, Grumbach K, Wang F, Fernandez A. Effects of limited English proficiency and physician language on health care comprehension. J Gen Intern Med. 2005;20(9):800–6.
5. Baker DW, Parker RM, Williams MV, Coates WC, Pitkin K. Use and effectiveness of interpreters in an emergency department. JAMA. 1996;275(10):783–8.
6. Sarver J, Baker DW. Effect of language barriers on follow-up appointments after an emergency department visit. J Gen Intern Med. 2000;15(4):256–64.
7. Schenker Y, Wang F, Selig SJ, Ng R, Fernandez A. The impact of language barriers on documentation of informed consent at a hospital with on-site interpreter services. J Gen Intern Med. 2007;22 Suppl 2:294–9.
8. Weech-Maldonado R, Morales LS, Spritzer K, Elliott M, Hays RD. Racial and ethnic differences in parents' assessments of pediatric care in Medicaid managed care. Health Serv Res. 2001;36(3):575–94.
9. Weech-Maldonado R, Fongwa MN, Gutierrez P, Hays RD. Language and regional differences in evaluations of Medicare managed care by Hispanics. Health Serv Res. 2008;43(2):552–68.
10. Weech-Maldonado R, Morales LS, Elliott M, Spritzer K, Marshall G, Hays RD. Race/ethnicity, language, and patients' assessments of care in Medicaid managed care. Health Serv Res. 2003;38(3):789–808.
11. Schenker Y, Karter AJ, Schillinger D, Warton EM, Adler NE, Moffet HH, Ahmed AT, Fernandez A. The impact of limited English proficiency and physician language concordance on reports of clinical interactions among patients with diabetes: the DISTANCE study. Patient Educ Couns. 2010;81(2):222–8.
12. Fernandez A, Schillinger D, Warton EM, Adler N, Moffet HH, Schenker Y, Salgado MV, Ahmed A, Karter AJ. Language barriers, physician-patient language concordance, and glycemic control among insured Latinos with diabetes: the Diabetes Study of Northern California (DISTANCE). J Gen Intern Med. 2011;26(2):170–6.
13. National Association of Community Health Centers. Serving patients with limited English proficiency: results of a community health center survey. June 2008. http://www.nachc.com/client/documents/LEP_report.pdf. Accessed December 1, 2011.
14. National Association of Community Health Centers. Community Health Centers: the local prescription for better quality and lower costs. March 2011. http://www.nachc.org/client/A%20Local%20Prescription%20Final%20brief%203%2022%2011.pdf. Accessed December 1, 2011.
15. National Association of Community Health Centers. About our health centers. http://www.nachc.org/about-our-health-centers.cfm. Accessed December 1, 2011.
16. National Association of Community Health Centers. Patients with limited English proficiency. Results of a community health center survey. http://bphc.hrsa.gov/uds. June 16, 2008. Accessed December 1, 2011.
17. Phokeo V, Hyman I. Provision of pharmaceutical care to patients with limited English proficiency. Am J Health Syst Pharm. 2007;64(4):423–9.

18. Dower C, McRee T, Grumbach K, Briggance B, Mutha S, Coffman J, Vranizan K, Bindman A, O'Neil EH. The practice of medicine in California: a profile of the physician workforce. California Workforce Initiative at the UCSF Center for Health Professionals, February 2001. http://www.futurehealth.ucsf.edu/Content/29/200102_The_Practice_of_Medicine_in_California_A_Profile_of_the_Physician_Workforce_executive_summary.pdf. Accessed December 1, 2011.

19. Taylor DA, Patton JM. The pharmacy student population: applications received 2008-09, degrees conferred 2008-09, fall 2009 enrollments. Am J Pharm Educ. 2010;74(10):S2.

20. Chisholm-Burns MA, Graff Zivin JS, Lee JK, Spivey CA, Slack M, Herrier RN, Hall-Lipsy E, Abraham I, Palmer J. Economic effects of pharmacists on health outcomes in the United States: a systematic review. Am J Health Syst Pharm. 2010;67(19):1624–34.

21. Chisholm-Burns MA, Kim Lee J, Spivey CA, Slack M, Herrier RN, Hall-Lipsy E, Graff Zivin J, Abraham I, Palmer J, Martin JR, Kramer SS, Wunz T. US pharmacists' effect as team members on patient care: systematic review and meta-analyses. Med Care. 2010;48(10):923–33.

22. Perez A, Doloresco F, Hoffman JM, Meek PD, Touchette DR, Vermeulen LC, Schumock GT; American College of Clinical Pharmacy. ACCP: economic evaluations of clinical pharmacy services: 2001-2005. Pharmacotherapy. 2009;29(1):128.

23. Laws MB, Heckscher R, Mayo SJ, Li W, Wilson IB. A new method for evaluating the quality of medical interpretation. Med Care. 2004;42(1):71–80.

24. Jurafsky D, Martin J. Speech and language processing: an introduction to natural language processing, computation linguistics, and speech recognition. 2nd ed. Upper Saddle River; Prentice Hall: 2008.

25. Williams, J. R. Guidelines for the use of multimedia in instruction, Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting. 1998. 1447–51.

26. McGuire LC. Remembering what the doctor said: organization and adults' memory for medical information. Exp Aging Res. 1996;22(4):403–28.

27. Kessels RP. Patients' memory for medical information. J R Soc Med. 2003;96(5):219–22.

Appendix A

ID	Type of concept	MAP recommendations
1	Medication side effect	Know the important side effects of simvastatin.
2	Medication side effect	These include unusual muscle pain or weakness that is not related to exercise.
3	Glycemic control	Test your blood sugar twice a day.
4	Glycemic control	When you wake up.
5	Glycemic control	One to two hours after lunch.
6	Medication change	You're going to use Novolin N 54 U in the morning and 44 U at night.
7	Medication administration	Take your glipizide with breakfast and dinner.
8	Medication adherence	Do not stop taking your losartan on your own.
9	Labs	You need to get your labs done next week.
10	Labs	Your kidney function needs to be retested.
11	Glycemic control	Take three or four glucose tablets for low blood sugar.
12	Medication education	Store your opened insulin vial at room temperature, not in the refrigerator.
13	Medication education	This will decrease the injection pain from insulin.
14	Medication education	You can write the one-month expiration date on it when you open it.
15	Medication education	Store all new vials of unopened insulin in the refrigerator.
16	Lifestyle	When you plan your meals, consider using the plate method three days a week—such as Monday, Wednesday, and Friday.
17	Glycemic control	Eat nuts rather than fruit for snacks in the evening.
18	Lifestyle	Let's increase your physical activity to 30 min a day three times a week.
19	Medication education	Avoid using salt substitutes when taking benazepril.
20	Medication adherence	Start aspirin again every day.
21	Referral	Make an appointment with your primary care doctor for your [condition]. (Rotate choice of following conditions: low back pain, urination pain, Viagra prescription, foot exam, annual physical checkup, Neurontin dosing.)
23	Glycemic control	Test your blood sugar twice a day, when you wake up and 1 to 2 h after lunch.
24	Glycemic control	Test your blood sugar twice a day, once before breakfast and once before dinner.
25	Glycemic control	Test your blood sugar twice a day, 1 to 2 h after breakfast and the other 1 to 2 h after dinner.
26	Glycemic control	Test your blood sugar twice a day, just before dinner and at bedtime.
27	Glycemic control	Test your blood sugar twice a day, when you first wake up and at bedtime.
28	Glycemic control	Test your blood sugar twice a day, before you exercise and after you exercise.
29	Glycemic control	Test your blood sugar twice a day, more often when you are sick.
30	Glycemic control	Test your blood sugar four times a day, once when you first wake up and 1 to 2 h after each of your three meals.
31	Glycemic control	Test your blood sugar once a day in the morning.
32	Glycemic control	Test your blood sugar at least once a week.
33	Glycemic control	Test your blood sugar when you have symptoms of low blood sugar.
34	Glycemic control	Test your blood sugar when you have symptoms of high blood sugar.