INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, 16(2), 141–161

The Development and Evaluation of a Speech-to-Sign Translation System to Assist Transactions

Stephen Cox
Michael Lincoln
Judy Tryggvason
School of Information Systems,
University of East Anglia

Melanie Nakisa
Royal National Institute for Deaf People

Mark Wells
Marcus Tutt
Sanja Abbott
Televirtual Ltd., Norwich

We describe the design, development, and evaluation of an experimental translation system that aims to aid transactions between a deaf person and a clerk in a post office (PO). The system uses a speech recognizer to recognize speech from a PO clerk and then synthesizes recognized phrases in British Sign language (BSL) using a specially developed avatar. Our main objective in developing this prototype system was to determine how useful it would be to a customer whose first language was BSL and to discover what areas of the system required more research and development to make it more effective. The system was evaluated by 6 prelingually profoundly deaf people and 3 PO clerks. Deaf users and PO clerks were supportive of the system, but the former group required a higher quality of signing from the avatar and the latter a system that was less constrained in the phrases it could recognize; both these areas are being addressed in the next phase of development.

The development of TESSA was sponsored by the United Kingdom Post Office and the evaluation by the European Fifth Framework Programme under the auspices of the ViSiCAST project. We thank the deaf participants, Post Office clerks, and interpreters for their help in the evaluations.

Request for reprints should be sent to Stephen Cox, School of Information Systems, University of East Anglia, Norwich NR4 7TJ United Kingdom. E-mail: sjc@sys.uea.ac.uk

2. OVERVIEW OF THE SYSTEM

2.1. Design Philosophy

Our goal was to develop a system to enable a PO counter clerk to communicate with a deaf customer using automatically generated sign language, and hence to aid completion of a transaction. A priori, it might seem that recognizing the clerk's speech and displaying it as text to the deaf customer would be adequate. However, for many people who have been profoundly deaf from a young age, signing is their first language; therefore, they learn to read and write English as a second language (Conrad, 1979). As a result, many deaf people have below-average reading abilities for English text and prefer to communicate using sign language (Wood, Wood, Griffiths, & Howard, 1986).

Having previously developed a prototype system (SignAnim; described in Bangham et al., 2000; Pezeshkpour, Marshall, Elliott, & Bangham, 1999) that used an avatar to provide signing of subtitles for television, an avatar system was already available that could be employed to produce signs. A problem with SignAnim, and also for developing the system reported in this article, was translation from text to BSL. Whereas systems to translate text from one spoken language to another are now available and work well within a restricted domain of discourse, translation from text to sign language is still a formidable research problem. BSL is a fully developed language, largely independent of English, with its own signs to express distinct concepts and with its own syntactic and semantic structures (Brien, 1992). These structures, inherent to sign languages, differ somewhat from those found in spoken languages, and hence, translation from text to sign language requires a different approach from the techniques used in automatic translation of spoken languages.

SignAnim circumvented the translation problem by translating subtitles into Sign Supported English (SSE) rather than BSL. SSE uses the same (or very similar) signs for words as BSL but uses English language word order. Thus, the SSE equivalent of "The man is standing on the bridge" is MAN + STAND + ON-BRIDGE, and for "The cat jumps on the ball" it is CAT + JUMP + ONTO + BALL. SSE may therefore be regarded as more like a system for "encoding" English. Linguists regard SSE as English translated into signs and don't consider it a language per se. SignAnim was an important starting point for the system described here: By bypassing many of the difficult problems of translation from English to sign language, it provided an opportunity to develop reliable sign capture methods, to determine how legible a virtual human signer could be, and to develop a real-time signing "engine" that integrated the whole system.

Using prestored SSE "words" enables sentences to be translated into sign language at the expense of using a language that is less acceptable than BSL to deaf people. An alternative approach is to use whole phrase units rather than words. This approach is possible only if a small number of phrases are required, and these phrases can be recorded in BSL rather than SSE. If recording of the signs is done correctly, phrases can be concatenated to a certain extent, e. g., amounts of money can be slotted into a carrier phrase such as "The cost is ...." Although this approach im-

poses considerable restrictions on the meanings that can be conveyed in BSL and hence on the dialogue, we considered that the limited nature of the transactions in the PO should mean that most transactions could be completed in this way. Furthermore, the philosophy of using prestored phrases enables the speech recognition to be implemented as a finite state network, which as has already been noted increases the accuracy of the system (see Section 2.2). It was important to see how far a BSL system using prestored phrases could be taken as the first step toward developing a more general system.

2.2. System Components

Figure 1 is a diagram showing the structure of the system.

The PO clerk wears a headset microphone. In early versions of the system, the clerk operated a "push-to-talk" switch when he or she wished to communicate with the deaf customer, but in later versions, the speech recognizer was constantly active and would respond when the clerk uttered a "legal" phrase from the grammar. The screen in front of the clerk displays a menu of topics available, e. g., "Postage," "DVLA," "Bill Payments," "Passports." Speaking any of these words invokes another screen showing a list of phrases relevant to this category that can be recognized. However, this is only an aide-mémoire to the clerk; all phrases are active (i.e., can be recognized) at any time, so that switching between categories is seamless. In trials, we found that the clerk could remember many of the most commonly used phrases without consulting the screen.

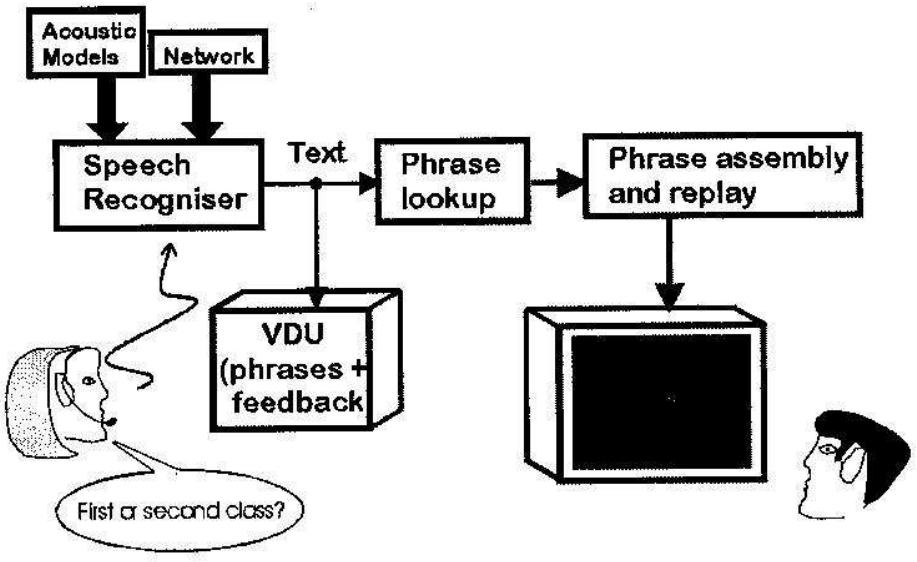

FIGURE 1 The post office translation system.

Prior to designing the system, we obtained transcripts of recordings of PO transactions at three locations in the UK, in all about 16 hr of business. Inevitably, much of the dialogue transcribed was in the nature of social interaction and had little to do directly with the transaction at hand. However, analysis of these transcriptions was essential for estimating the vocabulary that would be needed by the system to achieve a reasonable coverage of the most popular transactions. At the end of this analysis, a set of 115 phrases was prepared, which we estimated should be adequate to cover about 90% of transactions performed. This set of phrases was changed and extended after trials with users (see Section 3) and the total number of phrases currently available in the system is about 350.

2.3. Speech Recognition

In the first version of the system, the speech recognizer used was the Entropic HAPI (Hidden Markov Model Toolkit [HTK] Application Interface) system (Odell, Ollason, Valtchev, & Whitehouse, 1997), which incorporates the HTK recognizer (Jansen, Odell, Ollason, & Woodland, 1996). This had the advantage that it allowed us to experiment with using acoustic models that had been prepared in our own laboratory using the HTK software. The second version used the IBM ViaVoice recognizer (MODEL NUMBER, SUPPLIER NAME AND LOCATION), which offered greater range and flexibility in its network definition and in its user interface.

Both speech recognizers use the same underlying system: The speech signal is first parameterized into a sequence of vectors, each of which is formed from a 20 to 30 msec segment of the signal and extracts important information about this segment. The recognizer has stored speech models of several thousand *triphones* (phonemes in left and right context), each model consisting of a hidden Markov model (Cox, 1990) with a multivariate Gaussian mixture distribution of vectors associated with each state. A network of legal phrases is supplied to the recognizer, which uses a dictionary to rewrite each word within a phrase as a sequence of triphones. Decoding of the speech signal is done using an algorithm that uses the speech models and the network supplied to output the most likely sequence of words given the acoustic input and the network (for a detailed introduction to these topics that is relevant to the operation of the ViaVoice recognizer, see Jelinek, 1997).

An important point about the operation of the recognition system is that both the speech models and the network can be easily changed or adapted. The speech models can be adapted to the voice of each user ("speaker adaptation"), a process that takes about an hour, and the individual's models are then stored for later use. Speaker adaptation of the models greatly increases the recognition accuracy and hence the usability of the system. The fact that the network can easily be changed means that phrases can be altered or added to the system without the need for any recompilation.

The network constrains the speech recognizer to a finite number of predefined paths through the available vocabulary. These paths define the set of allowed phrases and consist of a start node (usually denoting silence, or background noise) followed by number of word nodes or subnetworks, finishing with an end node

(again denoting silence). Subnetworks are useful ways of defining phrase segments that can vary. For instance, a subnetwork called "one2hundred" represents the legal ways of saying the integers between 1 and 100, and this can be inserted at any appropriate point into the network. There are other subnetworks called "amounts-of-money," "days-of-the-week," "countries," etc. A fragment of the network is shown in Figure 2.

The use of a finite-state network may appear to place too much constraint on what can be said by the clerk. However, it is consistent with the philosophy outlined in Section 2.1 of using a limited set of prestored phrases for signing. Furthermore, once the clerk is familiar with the repertoire of phrases and the recognizer has been adapted to his or her voice, recognition performance is much higher than, for instance, currently available dictation packages. There are essentially two reasons for this:

1. Dictation packages are required to decode a large vocabulary and syntax and therefore use a probabilistic "bigram" language model in which the decoding of the speech utterance is controlled by the probability of any word in the vocabulary following any other word. Restricting the vocabulary and limiting the syntax to word sequences allowed by a network lowers the number of decoding possibilities very significantly and hence increases accuracy.
2. The recognizer can be operated on a "best-match" basis so that a phrase that is phonetically "close" but not identical with a phrase in the network will be recognized as the latter. This allows some flexibility for the speech of the clerk. (For instance, the phrase "Put that on the scales, please," which is not present in the network, would be recognized as "Please put it on the scales").

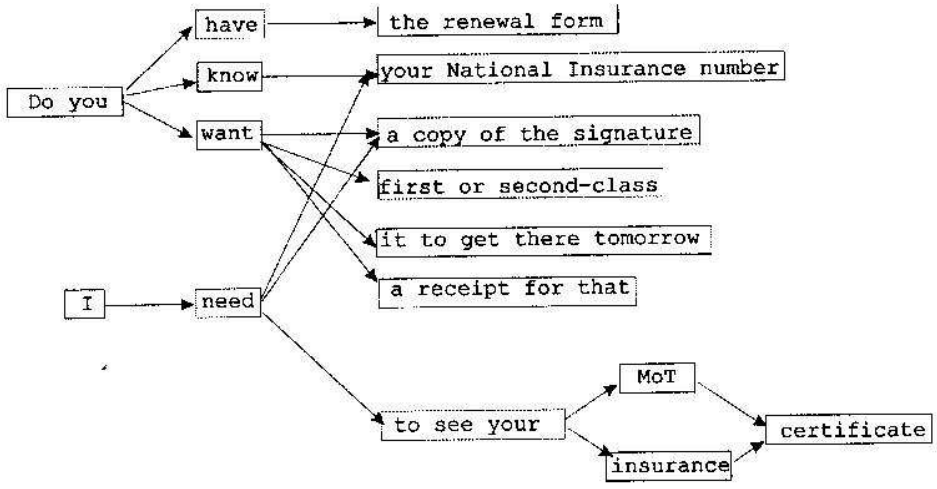

FIGURE 2 A section of the recognition network.

High recognition accuracy is very important for our system: The translation process is inherently slow because the avatar signs rather slowly to achieve maximum clarity, and any extra delay due to correcting mistakes made by the recognizer is likely to make the system unusable. Note also that because there is no separation of speech and language decoding in this system, it does not suffer from inaccuracies in the speech decoding process being forwarded to a language translation process that is also imperfect, an effect that can make more complex systems fail to translate correctly even quite simple phrases. By using prestored phrases, we in effect trade flexibility and range for accuracy.

The system described here is the first stage toward a more sophisticated system that will incorporate the techniques used in "speech-understanding" systems to enable a much wider range of transactions to be completed. In this research system, we are experimenting with using a probabilistic language model recognizer followed by a language processor that attempts to map the output from the recognizer to the correct phrase. This has the benefit of allowing the clerk complete flexibility in what he or she says to the recognizer (as long as the words used are within the approximately 100,000-word vocabulary of the recognizer) at the expense of requiring some language understanding to determine the correct sequence of signs to be output. At the time of writing this article, we do not know whether or not this system will be less accurate than the system that uses a network. In addition, the system can obviously be adapted to translate to another spoken language (using either displayed text or speech output) as well as to sign language, and this possibility is also being explored.

2.4. System Software

The system software has the task of enabling communication between the speech recognition module and the avatar module and of controlling the overall progress of a transaction. The sign assembly system is written in TCL and the recognition module incorporated as a TCL extension. The avatar module is written in C++, and communication between this and the other system components is performed using a remote procedure call system via TCP/IP socket connections.

2.5. The Signing Avatar, TESSA

The simplest way of signing the set of phrases defined for the application would be to store video recordings of a person signing each phrase and concatenate the appropriate phrases in response to the output from the speech recognizer. However, we have been developing an experimental system that uses a virtual human (avatar) to sign Teletext subtitles (Wells, Pezeshkpour, Tutt, Bangham, & Marshall, 1999; Teletext, developed in the 1970s by the British Broadcasting Corporation, consists of pages of information such as news and sports that are viewed on a television set capable of viewing these pages). In this broadcast application, using an avatar has an important advantage over using video in that the signing can be transmitted us-

ing a very small bandwidth (only the model positions need to be transmitted at suitable intervals rather than a full video signal). Although bandwidth is not a consideration for the PO system described here, an ultimate aim within the ViSiCAST project is to produce a "text-to-sign" synthesizer that will be capable of synthesizing signs from a much less restricted vocabulary; to build such a system using concatenated video clips would not be viable. Another advantage of using an avatar is that different figures can be rendered onto the avatar's frame so that a single set of recordings of signs can be used to drive different virtual humans. Conversely, multiple human signers can be used to generate the signed content of the system while using the same avatar for the output signing, making it easy to expand and update the signed content. In addition, concatenation of signing is more fluent and controlled for avatar than for video signing, as the exact positioning of the avatar can be manipulated. For these reasons, we decided to disply the signs using an avatar, TESSA, which was based on the avatar used in the SignAnim project.

Research into methods for capturing signing movements directly from video has been reported (Ahmad, Taylor, Lanitis, & Cootes, 1997; Huang & Huang, 1998; Lien & Huang, 1998; Starner, Weaver, & Pentland, 1998). This approach is highly desirable, as it obviates the need to record signs by attaching motion sensors to a human with the attendant problems of invasiveness, motion restriction, calibration, sensor fusion, and so forth. Unfortunately, capture from video is not yet robust enough to record high-quality motion. The alternative is to capture signs using separate sensors for the hands, body, and face. This technique appears to capture sufficient movement to generate true and realistic signing from a virtual human.

The motion is captured as follows:

1. Cybergloves with 18 resistive elements for each hand are used to record finger and thumb positions relative to the hand itself.
2. Polhemus magnetic sensors record the wrist, upper arm, head, and upper torso positions in three-dimensional (3D) space relative to a magnetic field source.
3. Facial movements are captured using a helmet-mounted camera with infrared filters and surrounded by infrared light emitting diodes to illuminate Scotchlight reflectors stuck onto the face. Typically, 18 reflectors are placed in regions of interest such as the mouth and eyebrows.

Figure 3 shows this configuration in use.

The sensors are sampled at between 30 and 60 Hz and the separate streams integrated, using interpolation where necessary, into a single, raw motion-data stream that can drive the virtual human directly. The system is calibrated at the beginning of each session, but in practice, the main variation lies between signers. For example, the considerable cross talk between glove sensors depends heavily on how tightly the gloves fit. It is particularly important to ensure good calibration at positions where fingers are supposed to just touch the thumb and where hands touch both each other and the face. These positions are important to clear signing and, to reduce computation times, there is currently no collision detection to prevent body parts sinking into each other. Where individual signs or segments are to be added

FIGURE 3 Data capture: Face tracking camera with facial reflectors, Cybergloves for tracking the digits, and Polhemus sensors taped onto the back of each hand and upper arm, the body, and the head to track the body.

to the lexicon, then signs are altered manually, using a custom editor program, and the beginning and end of each sign is marked to aid concatenation.

The motion-data stream is displayed using a virtual human. In common with many avatars, a 3D "skeleton" is driven directly from the motion data. The skeleton is wrapped in and elastically attached to a texture mapped, 3D polygon mesh that is

controlled by a separate thread (event loop) that tracks the skeleton. One of the latest PC-accelerated 3D graphics cards is used to render the resulting 5,000 polygons at 50 frames per second using Direct–X on a Pentium class PC. Because TESSA is a full 3D model, her position and pose can be changed by the user during use, an extremely valuable feature that enables users to select the optimal viewing angle and size. In addition, the identity of the virtual human can be changed. TESSA is capable of signing in real time with a refresh rate of approximately 40 frames per second.

Figure 4 gives an idea of the appearance of the avatar as it makes BSL signs for some of the days of the week. This avatar was based on a mesh "library," whereas the later version was based on a 3D scan of a human participant and is more lifelike.

3. EVALUATION

It is essential that the system conveys useful information in a way that is helpful and acceptable to deaf users. The extent to which TESSA met this aim was assessed by the measurement of three areas of performance:

1. The quality of the signs.
2. The difficulty of performing a transaction with TESSA.
3. The perceptions of the deaf users and the PO clerks of the system.
4. The outcomes of these experiments are reported in this section.

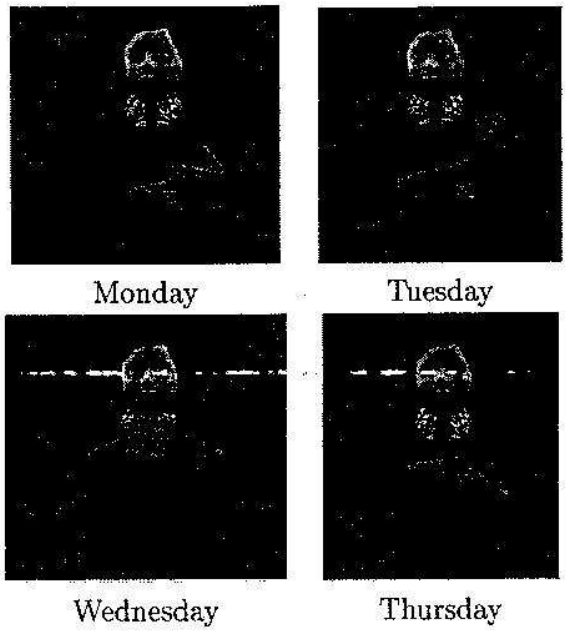

FIGURE 4 Stills from the signs for the four days: Monday, Tuesday, Wednesday, and Thursday.

3.1. Participants and Protocol

Six prelingually profoundly deaf people whose first language is BSL took part in the evaluations of the system. They were recruited through the deaf-UK e-mail newsgroup or through local UK Royal National Institute for Deaf People offices and were paid for their participation. Three clerks were recruited by the PO to take part in the evaluations: Each had over 10 years experience as a clerk and had experience of serving deaf customers.

The evaluations took place over three sets of 2 days. Two deaf people and one clerk attended for each pair of days. The 1st day started with completion of the first part of a questionnaire. Each deaf participant then alternated between identifying a block of signed phrases and attempting a block of staged transactions. At the end of the 2nd day, all participants completed the remainder of the questionnaire and gave any general feedback. BSL/English interpreters were present throughout.

3.2. Experimental Design and Procedures

Measurement of sign intelligibility and acceptability. The quality of TESSA's signing was measured in two ways: intelligibility of signs and acceptability of signs to deaf users. The first of these measurements is an objective one and is clearly important in establishing a baseline for this system against which future avatars may be evaluated. However, it is well-known that intelligibility on its own is inadequate for assessment of these systems: for instance, synthetic speech can sound fully intelligible but be disliked by users (Johnston, 1996). Hence, we also included a subjective measurement of acceptability of signing.

The deaf participants were presented with each signed phrase and asked to write down what they understood. From the 115 distinct phrases, 133 phrases were generated by incorporating days of the week and numbers to ensure that each day and each number (units and 10s) was presented at least once. Signed phrases were presented on the screen without text. The deaf participant controlled presentation of each phrase and was allowed to repeat each phrase up to a maximum of five presentations. Phrases were presented in blocks of between 20 and 24 in groups according to broad categories, for example, postage, bill payment, and amounts of money. Accuracy of identification of phrases was assessed in two ways:

1. By the accuracy of identification of complete phrases.
2. By the accuracy of approximate "semantic sign units" within the phrase. For example, the phrase "It should arrive by Tuesday but it's not guaranteed" requires five sign units; therefore, "should arrive Tuesday not guaranteed" would score 100% and "should arrive Tuesday" 66%.

The 133 phrases gave a total of 444 sign units. Although these units were not all distinct (e.g., the sign for "pound" was presented several times), identification of each presentation of a unit was scored separately. One experimenter (Judy Tryggvason) judged the accuracy of responses for both measures on the basis of

written responses from each deaf participant. Once each phrase had been scored for accuracy of identification, each deaf person was re-presented with each phrase not identified correctly along with the text of the intended phrase. With an interpreter and experimenter, they were asked to indicate whether the signs were considered inappropriate or whether they were just not clear. Any signs considered inappropriate were not necessarily wrong; rather, they may have represented different regional variations in sign to those used by the deaf participant. (Variation in signs is a more difficult problem to contend with than variations in accent or dialect in spoken languages, as hearing people can use a standard written language as a reference, which is not available to those who communicate using only signs; Kyle & Woll, 1985.)

Participants were also asked to rate how acceptable the phrase was as an example of BSL on a 3-point scale ranging from 1 (*Low*) to 3 (*High*).

Measurement of effectiveness of TESSA in transactions. Staged PO transactions were used to compare completion times and ease and acceptability of communication with and without TESSA. Each deaf participant attempted 30 transactions with a single PO clerk. Transactions were selected by the PO as those achievable with the phrases available. There were 18 distinct transactions; 6 were denoted "simple," 6 "average difficulty," and 6 "complex." The average difficulty and complex transactions were attempted twice by each deaf participant and clerk pair, once with an open counter and once behind a fortified counter where a transparent screen separates clerk and customer. Use of different counter styles did not appear to affect performance; hence, results are not reported separately here.

Half of all transactions were attempted with TESSA and half without. The phrases presented with or without TESSA were counterbalanced between deaf participants. Practice transactions were performed with TESSA at the start of each session so that the clerk, deaf participant, and interpreter could get used to using TESSA and the format of the evaluation. Transactions were performed in blocks of six, three with TESSA and three without. The approximate time taken to successfully complete each transaction was recorded. On completion of each transaction, both deaf participants and clerks were asked to rate each transaction for acceptability on a 3-point scale ranging from 1 (*Low*) to 3 (*High*).

Measurement of subjective opinions about the system. Questionnaires to both deaf participants and clerks were used to obtain subjective views of previous experiences of communication in the PO, and how these experiences differed in the trials and were anticipated to differ in real life using TESSA.

3.3. Results

Quality of signing. The average number of times each phrase was presented before an attempt at identification was made was 1.8. Attempts at identification

were made after one presentation for the majority of phrases (51%) and required more than two presentations for 20% of the phrases. The average accuracy of identification of complete phrases was 61% and ranged from 42% to 70% across deaf participants (Figure 5). For the identification of sign units in phrases, average accuracy was 81% and ranged from 67% to 89% (Figure 5).

Subsequent analysis of the sign units that were wrongly identified indicated that on average 30% of errors (6% of all sign units) were due to signs considered inappropriate and the remaining 70% (13% of all sign units) were due to unclear signing.

Table 1 shows the percentage of phrases that were rated in each category of acceptability. The average acceptability rating was 2.2 and ranged from 1.7 to 2.8.

Transactions. On average, transactions took longer to complete with TESSA than without, $F(1, 178) = 61.2$, $p < .001$ (Figure 6).

Average times for transactions were 57 sec without TESSA and 112 sec with TESSA. On average, communication in transactions completed with TESSA was rated by deaf participants as less acceptable than in transactions completed without TESSA, $U(1, 178) = 6025$, $p < .001$ (Figure 7).

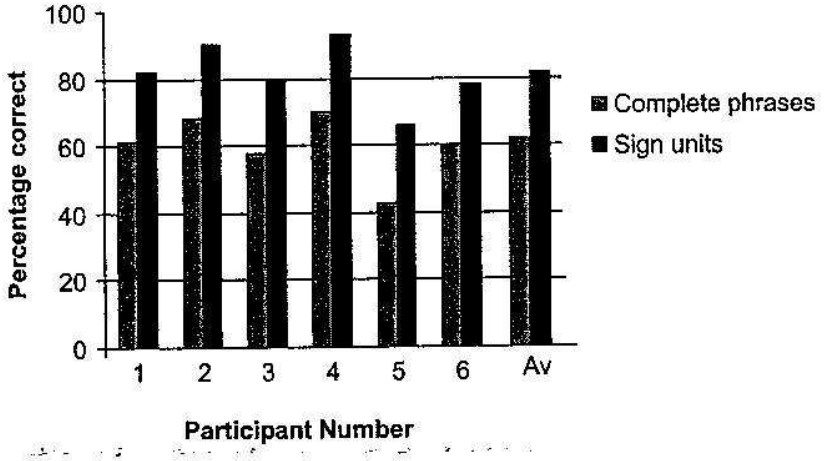

FIGURE 5 Average percentage recognition scores achieved by each signer for complete phrases and sign units within phrases.

Table 1: The Percentage of Phrases Rated in Each Category of Acceptability of Signing Quality

Acceptability Rating		*% of Phrases*
High	3	20.2
	2	43.3
Low	1	36.6

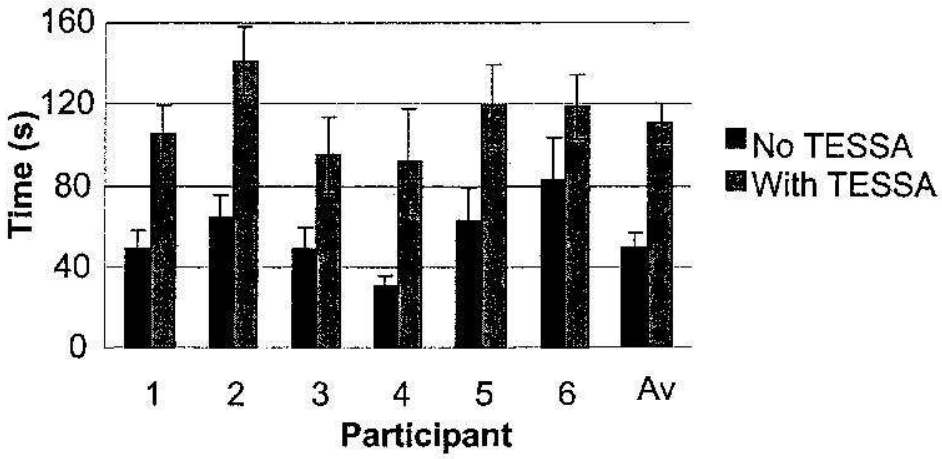

FIGURE 6 Average times of translations without Tessa (dark bars) and with TESSA (light bars) for each deaf participant. Error bars show the 95% confidence intervals of the means.

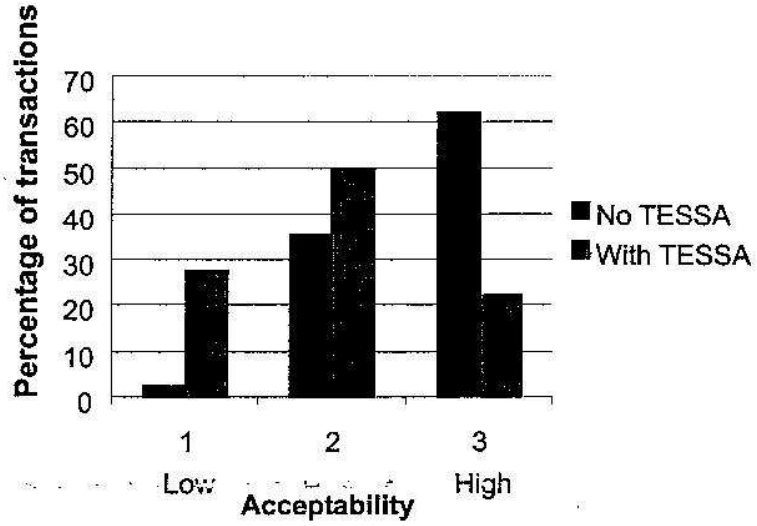

FIGURE 7 Percentage of transactions rated by the deaf participants in each category of acceptability on a 3-point scale ranging from 1 (*Low*) to 3 (*High*) without TESSA (dark bars) and with TESSA (light bars).

On the 3-point scale ranging from 1 (*Low*) to 3 (*High*), average ratings of transaction acceptability were 1.9 with TESSA and 2.6 without. Clerks rated acceptability of transactions completed with TESSA as slightly lower than transactions completed without TESSA. On the 3-point scale, average ratings were 2.5 with TESSA and 2.6 without (Figure 8).

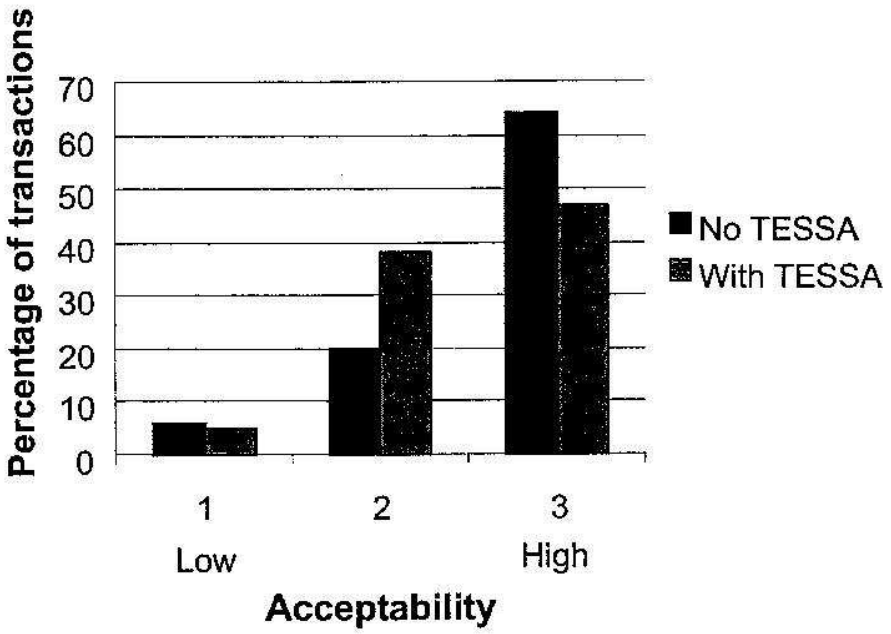

FIGURE 8 Percentage of transactions rated by the clerks in each category of acceptability on a 3-point scale ranging from 1 (*Low*) to 3 (*High*) without TESSA (dark bars) and with TESSA (light bars).

Subjective opinions. The deaf participants were asked three questions about ease of communication in the PO, namely, how easy they normally found communication without TESSA, how easy it was in these trials with TESSA, and how easy they anticipated it would be in everyday life with TESSA. They were asked to rate their ease of communication on a 5-point scale ranging from 1 (*Very difficult*) to 5 (*Very easy*). The questions, with the mean responses, standard deviations, and range of responses, are shown in the first three rows of Table 2. The participants were then asked an additional two questions about the extent to which communication in the PO upset them, and they responded on a 5-point scale ranging from 1 (*Very much*) to 5 (*Not at all*). The questions, with the mean responses, standard deviations, and range of responses are shown in rows 4 and 5 of Table 2.

The PO clerks were questioned about communicating with deaf people in the PO previously, with TESSA in the trials, and with TESSA in everyday life, rated on a response scale ranging from 1 (*Very difficult*) to 5 (*Very easy*). Rows 6, 7, and 8 of Table 2 show the questions together with the mean responses, standard deviations, and ranges. All clerks said that they would prefer to have TESSA available as an option to use when communication became difficult, even though they all thought transactions would take "Slightly longer" with TESSA. The clerks were then asked two questions about whether TESSA made communication with a deaf customer easier, rated on a response scale ranging from 5 (*Much easier*) to 1 (*Much worse*). The questions, mean responses, standard deviations, and ranges are shown in rows 9

Table 2: Summary of Responses to Questions to Deaf Participants and to Post Office Clerks About the TESSA System

Question	*Response Scale*	*M Response*	*SD*	*Range*
How easy do you usually find communication in the Post Office?	1 = *Very difficult* 2 = *Slightly difficult*	2.7	1.5	1 to 5
How easy did you find communication using TESSA?	3 = *Manageable* 4 = *Fairly Easy*	2.5	0.83	1 to 3
In everyday life, how easy do you think communication would be using TESSA?	5 = *Very Easy*	2.5	1.37	1 to 4
In everyday life, how much does communication in the Post Office upset, annoy, or worry you?	1 = *Very much* 2 = *Quite a lot*	3.33	1.63	1 to 5
In everyday life, how much would communication using TESSA in the Post Office upset, annoy, or worry you?	3 = *Some* 4 = *A little* 5 = *Not at all*	4.33	1.21	2 to 5
How easy do you usually find communication with deaf customers?	1 = *Very difficult* 2 = *Slightly difficult*	4.0	0.0	4 to 4
How easy did you find communication using TESSA?	3 = *Manageable* 4 = *Fairly easy*	4.33	0.57	4 to 5
In everyday life, how easy do you think communication would be using TESSA?	5 = *Very easy*	4.66	0.57	4 to 5
Compared to communication without, do you think TESSA made communication?	1 = *Much worse* 2 = *Slightly worse*	4.33	0.57	4 to 5
In everyday life, do you think that using TESSA in the Post Office would make communication?	3 = *No difference* 4 = *Slightly easier* 5 = *Much easier*	5.0	0.0	5 to 5

and 10 of Table 2. All clerks said communication was "Slightly easier" or "Much easier" with TESSA than without and were unanimous that in everyday life they expected that communication would be "Much easier" with TESSA.

3.4. Discussion

Intelligibility and acceptability of signing. Accuracy of identification of the signed phrases was 61% for complete phrases and 81% for sign units, with quite a wide range in accuracy across deaf participants (ranges of 28% and 20%, respectively). This range in accuracy suggests it is important to use many sign language users for a true assessment of signed content of these systems. In the future, it may be more appropriate to use more than six deaf people from a range of UK regions to assess sign quality.

The majority of identification errors (70%) were due to signs being unclear rather than due to inappropriate signs. The percentage of errors for inappropriate signs did not differ greatly between participants, with personal averages ranging from 4.7% to 6.6%. This pattern might suggest that the same signs were considered inappropriate

by all deaf participants. However, inspection of the pattern of errors across deaf participants for each phrase indicated that this was not necessarily the case. Of the 46 phrases in which one or more sign was considered inappropriate by any deaf participant, in 34 of these (74%) a sign was considered inappropriate by no more than two of the deaf participants. This result suggests that regional variations or differences in personal signing style may have played a role in phrase intelligibility.

Ratings of acceptability were also given across the scale with 20% of phrases rated as highly acceptable and 63% in one of the top two categories, indicating that there is scope for improving the quality of the avatar's signing.

Transactions using TESSA. Compared to transactions without TESSA, transactions performed with TESSA took on average nearly twice as long to complete, and the deaf participants, and to a lesser extent the clerks, rated communication as less acceptable. The main reason most likely to have contributed to these effects was the somewhat disjointed communication with TESSA. As expected, it took the clerks some time to learn which phrases were available and to locate the phrase they wanted so they could read it out word for word. The clerks had only about an hour of practice using the system before the trials. These difficulties should decrease substantially with training and experience on the system. Moreover, the next version of the system, which will incorporate some speech understanding, will not require phrases to be repeated verbatim.

Additional factors may have contributed to the longer transaction times and poorer ratings with TESSA:

1. The list of phrases was selected for use in the system as those most commonly used in the PO. These phrases also tended to be those used for the more simple PO transactions, for example, buying stamps, cashing a check, or claiming a pension payment. Hence, the transactions used in this evaluation, limited by the phrases available, also tended to be fairly simple or were simplified. This was confirmed by the PO staff who selected the transactions and the clerks who said they would usually ask more questions for specific transactions but these were not available in TESSA. The transactions used in the trials therefore tended to represent situations in which communication was fairly easy without TESSA.
2. The deaf participants were all fairly good communicators and all had reasonable written skills. Hence, they were able to complete the simple transactions by lip reading/speaking and writing notes or asking the clerk to write things down when necessary. This is a consequence of the type of people who would be prepared to attend 2 days of testing away from their home town, the recruitment process (through e-mail and professional connections), and also the necessary use of textphone, fax, and e-mail for the logistics of arranging the trials.
3. The clerks either were "deaf aware" or soon became deaf aware as a result of spending 2 days with the profoundly deaf participants. Communication without TESSA was fairly easy, as they used good eye contact, spoke clearly, and were prepared to write things down if they were not understood.

4. There was a delay of a few seconds between recognition of the spoken phrase and the signing of the phrase. Not only did this absolute delay add to overall transaction time, but the delay often resulted in loss of attention and the need for the sign to be repeated or the clerk to repeat the phrase.

Questionnaires. The small sample size (six deaf participants and 3 PO clerks) makes interpretation of the questionnaires problematical. The responses to the questions posed to the deaf participants (shown as rows one to five of Table 2) do not allow any clear inferences to be drawn about the utility of the TESSA system to the deaf participants, apart from the fact that TESSA did not make communication significantly worse. The responses to the questions posed to the PO clerks (shown as rows six to ten of Table 2) show that the three clerks responded positively to TESSA, but no more than that. However, it does not seem unreasonable that the responses were not more generally positive at this stage in the life cycle of the project. The questions were asked about the first version of TESSA to be evaluated by deaf people and on the basis of use during the trials by clerks with little previous experience of using the system in which communication with TESSA was somewhat lengthy and disjointed.

The deaf participants provided much constructive feedback about how TESSA could be improved. Their main points were:

1. Facial expressions need to be improved.
2. Clearer hand shapes, finger configurations, and lip patterns are required, especially for numbers and finger spelling.
3. The delay between the end of the spoken phrase and the beginning of signing needs to be reduced.
4. The appearance of the avatar needs to improve. In particular, a clearer distinction should be made between the face and hands and the clothing, which should be plain.
5. All deaf participants said they would prefer to see both BSL and text rather than just BSL or just text. They also thought that SSE should be available as an option.

When asked to comment on the use of avatars for signing in general, all deaf participants thought that avatars would be most useful for more complex communication needs, for example, explaining forms to claim social benefits.

All clerks said they would prefer to have the system available, as they thought it would make communication with deaf customers easier and more effective. Use of the system for multiple spoken languages and with text subtitles would ensure more frequent use and hence greater likelihood that the system would be used with deaf people. The clerks also commented that they would like more phrases and an unconstrained speech system in which phrases need not be spoken verbatim.

4. GENERAL COMMENTS AND FUTURE WORK

The goal in developing this trial system was to establish whether the introduction of a limited speech-to-sign translation system for the PO counter clerk would be beneficial to deaf users whose primary means of communication was sign language. Although some of the feedback from the evaluation was critical, we are encouraged by the positive comments received from both groups of participants in the trials who uniformly contended that an improved system would be very beneficial.

The evaluations, although limited in extent, have indicated that there is much scope for improvement of TESSA and of similar systems. They have given some insight into how these improvements could be achieved and provided baseline outcome measures against which improvements can be assessed. The majority of aspects identified for improvement are planned for further development within the ViSiCAST project. The most important of these is the development of an "unconstrained speech input" version in which phrases need not be repeated word for word by the clerks. This will reduce considerably the time taken for transactions and hence should make the system more acceptable to both deaf customers and clerks. Other aspects to be explored include research into facial modeling, which will improve avatar facial expressions and lip patterns. New data gloves are also being used to improve recording of finger movements and hand shapes. New models of the avatar and clothing will also take account of the comments made by the deaf participants.

In this article, we have reported the results of a first evaluation of the system carried out when it was at an early stage of development. Since then, a more advanced system has been trialled in a PO over a period of several weeks. Reports from the PO clerks who have been using the system indicate that as they have become more familiar and practiced with the system, they have gained in confidence, and their ability to use the system has increased. We do not yet have any feedback from deaf users who have used the system on several occasions over a period of time. However, the system is currently being trialled in five large UK POs. When these trials are over, we will have an opportunity to gauge the reaction of both clerks and deaf users who interact using the system regularly over a long period. Less formal evaluations are planned within the deaf community to assess the views of more deaf people, and further formal evaluations will continue through the lifetime of the ViSiCAST project.

In tandem with these developments, the ViSiCAST project has also been doing basic research into the general problem of converting arbitrary English text into a representation of sign language (Safar & Marshall, 2001) and developing a synthetic avatar that can sign these representations without the need for motion capture (Kennaway, in press). These will feed into the application described here to increase its flexibility and sophistication. The problem of two-way communication is also being addressed by research into sign-language recognition.

REFERENCES

Ahmad, T., Taylor, C., Lanitis, A., & Cootes, T. (1997). Tracking and recognizing hand gestures, using statistical shape models. *Image and Vision Computing, 15,* 345–352.

Bangham, J., Cox, S., Lincoln, M., Marshall, I., Tutt, M., & Wells, M. (2000). Signing for the deaf using virtual humans. In *Proceedings of the IEEE colloquium on speech and language processing for disabled and elderly people.*

Brien, D. (1992). *Dictionary of British sign language / English.* CITY, STATE (COUNTRY): Faber and Faber.

Conrad, R. (1979). *The deaf school child.* New York: Harper & Row.

Cox, S. (1990). Hidden Markov models for automatic speech recognition: Theory and application. In C. Wheddon & R. Linggard (Eds.), *Speech and language processing* (pp. 209–230). LOCATION: Chapman and Hall.

Huang, C., & Huang, W. (1998). Sign language recognition using model-based tracking and a 3D Hopfield neural network. *Machine Vision and Applications, 10,* 292–307.

Jansen, J., Odell, J., Ollason, D., & Woodland, P. (1996). *The HTK book.* CITY, STATE (COUNTRY): Entropic Research Laboratories Inc.

Jelinek, F. (1997). *Statistical methods for speech recognition.* Cambridge, MA: MIT Press.

Johnston, R. (1996). Beyond intelligibility—the performance of text-to-speech synthesisers. *BT Technology Journal, 14,* 100–110.

Johnston, R., et al. (1997). Current and experimental applications of speech technology for Telecom services in Europe. *Speech Communication,* 23, 5–16.

Kennaway, J. (in press). Synthetic animation of deaf signing gesture. In I. Wachsmuth (Ed.), *4th international workshop on gesture and sign language based human–computer interaction.* New York: Springer Verlag.

Koo, M., et al. (1995). KT–STS: A speech translation system for hotel reservation and a continuous speech recognition system for speech translation. In *Proceedings of 4th European Conference on Speech Communication and Technology* (pp. 1227–1230).

Kyle, J., & Woll, B. (1985). *Sign language: The study of deaf people and their language.* Cambridge, England: Cambridge University Press.

Lien, C. C., & Huang, C. L. (1998). Model-based articulated hand motion tracking for gesture recognition. *Image and Vision Computing, 16,* 121–134.

Mazor, B., & Zeigler, B. L. (1995). The design of speech-interactive dialogs for transaction-automation systems. *Speech Communication, 17,* 313–320.

Morimoto, T. (1993). ATR's speech translation system: ASURA. In *Proceedings of the 3rd European Conference on Speech Communication and Technology* (pp. 1291–1294).

Odell, J., Ollason, D., Valtchev, V., & Whitehouse, D. (1997). *The HAPI book.* CITY, STATE (COUNTRY): Entropic Cambridge Research Laboratory.

Pezeshkpour, F., Marshall, I., Elliott, R., & Bangham, J. (1999). Developing of a legible deaf signing virtual human. In *IEEE Multimedia Systems Conference '99 (IEEE ICMCS '99)* (pp. 333–338).

Rayner, M., Alshawi, H., Bretan, I., & Carter, D. (1994). A speech to speech translation system built from standard components. In *Proceedings of the ARPA Human Language Technology Workshop '93* (pp. 217–222). Princeton, NJ: PUBLISHER NAME.

Safar, E., & Marshall, I. (2001). The architecture of an English-text-to-sign-languages translation system. In G. Angleova et al. (Eds.), *Recent advances in natural language processing (RANLP)* (pp. 223–228). CITY, STATE (COUNTRY): PUBLISHER NAME.

Starner, T., Weaver, J., & Pentland, A. (1998). Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1371–1375.

Wahlster, W. (Ed.). (2000). *Verbmobil: Foundations of speech to speech translation.* New York: Springer-Verlag.

Waibel, A. (1996). Interactive translation of conversational speech. *Computer, 29,* (7).

Wells, M., Pezeshkpour, F., Tutt, M., Bangham, J., & Marshall, I. (1999). Simon—an innovative approach to deaf signing on television. In *Proceedings of the International Broadcasting Convention* (pp. 477–482).

Wood, D., Wood, H., Griffiths, A., & Howarth, I. (1986). *Teaching and talking with deaf children*. New York: Wiley.

Wyard, P., et al. (1998). Spoken language systems—Beyond prompt and response. In F. Westall, R. Johnston, & A. Lewis (Eds.), *Speech technology for communications* (pp. 487–520). CITY, STATE (COUNTRY): Chapman and Hall.

Yoshioka, O., Minami, Y., & Shikano, K. (1995). A speech dialogue system with multi modal interface for telephone directory assistance. *IEICE Transactions on Information and Systems, E78D*, 616–621.