THE DEVELOPMENT AND EVALUATION OF A SPEECH TO SIGN TRANSLATION SYSTEM TO ASSIST TRANSACTIONS

M. LINCOLN[1], S.J. COX[1], M. NAKISA[2]

ABSTRACT. We describe the design, implementation and testing of an experimental interactive translation system that aims to aid a deaf person in transactions in a Post Office (PO). The system uses a speech recogniser to recognise speech from the PO clerk and then synthesises the recognised phrase in British Sign language (BSL) using a specially developed avatar. Our concern was to determine whether such a system was useful to a customer whose first language was signing and to discover what areas of such a system required more research and development to make it more effective. The system was evaluated by six pre-lingually profoundly deaf people and three PO clerks over a period of six days. The evaluation revealed that the deaf users require a higher quality of signing from the avatar and the clerks a system that is less constrained in the phrases it can recognise: both these problems are being addressed in the next phase of the development.

1. INTRODUCTION

There has recently been considerable research activity in developing automatic systems which can understand and output speech to provide information services or to perform transactions with customers [12]. Most of these systems have been developed for use over the telephone network with the goal of replacing completely or assisting a human operator [8]. A key aspect of them is that they operate in a rather restricted domain of discourse (e.g. train timetable enquiries [10], e-mail access [20], directory enquiries [21]) and this gives them some robustness to the difficult problems of variability and "noise" in the language used by the speakers, the speech signal and the telephony channel. There has also been work on interactive speech-to-speech translation systems. These systems are designed to provide translation of conversational speech with a potentially very large vocabulary [18]. We have been developing a system which combines aspects of both the kinds of systems mentioned above. It is an interactive translation system but it operates in a very restricted domain and is designed to assist in the completion of a transaction between a Post Office clerk and a deaf or hearing-impaired customer. The system translates the clerk's speech into British Sign Language (BSL) and displays the signs using a specially developed avatar.

The fact that very many of the transactions in a Post Office are highly predictable in their scope and progress is very important, as it enables us to achieve a good success rate using a very limited recognition vocabulary and without using such techniques as semantic parsing, which attempts to extract the meaning from an utterance. However, the system described here is the first stage towards a more sophisticated system which will incorporate the techniques used in "speech understanding" systems to enable a much wider range of transactions to be completed. Our concern in developing the current system was to find out whether a simple translation system would be of any use at all to the deaf community and to learn what the most important problems are with such an approach.

The system has been developed as part of the European Union fifth framework project, ViSiCAST [1], which aims to benefit deaf citizens by allowing them access to information and services in sign language.

2. Overview of the system

2.1. **Design philosophy.** Our motivation was to develop a system to enable a Post Office counter-clerk to communicate with a deaf or hearing-impaired customer using automatically-generated sign-language, and hence to aid completion of a transaction. *A priori*, it might seem that recognising 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 tends to be their first language and they learn to read and write more slowly. As a result, numbers of deaf people have below average reading abilities for English text.

We had already developed a prototype system (SignAnim, described in [1, 14]) that used an avatar to provide signing of sub-titles for television and so we had available an avatar that could be controlled to produce signs. A problem with SignAnim, and also with the system that we were to develop, 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 English 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 semantical structures [5]. These structures are less well understood than those of spoken languages [6].

SignAnim circumvented the translation problem by translating sub-titles 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*. Translating English into SSE was an important first step in the development of SignAnim as it provided an opportunity to develop reliable sign capture methods, to determine how legible a virtual human could be and to develop a real time signing "engine" that integrated the whole system. However, for the deaf community, SSE is unquestionably less popular than BSL.

Another approach to the translation problem is to use pre-stored phrases and to pre-record the signs (as avatar movements) for these phrases. If only a small number of phrases is required, it is possible to record these in BSL rather than SSE. Also, 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 imposes considerable restrictions on the meanings that can be conveyed by the clerk and hence on the dialogue, it has the advantage that BSL can be used. Furthermore, we considered that the limited nature of the transactions in a Post Office should mean that most transactions could be completed in this way. In addition, it was important to see how far a system using pre-stored phrases could be taken as the first step towards developing a more general system. Using pre-stored phrases also confers benefits: the speech recognition is potentially very accurate because of the limited number of vocabulary items to be recognised and one can also be sure that the meaning of a phrase uttered is accurately translated into the target language. These gains are important ones, as the "noise" introduced into the

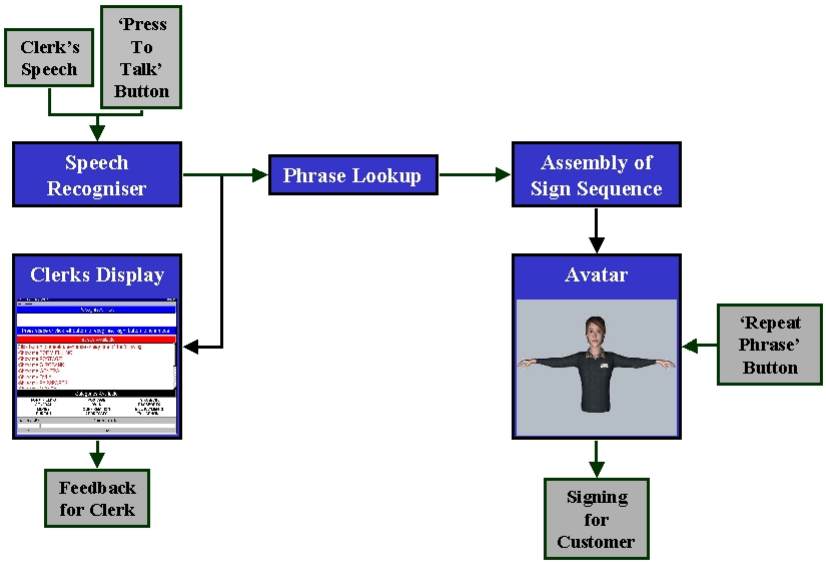

FIGURE 1. The Post Office translation system

information channel by inaccuracies in the recognition process combined with ambiguities in the translation process can make more complex systems fail to translate correctly even simple phrases. By using pre-stored phrases, we in effect trade flexibility for accuracy.

2.2. **System components.** Figure 1 is a diagram showing the structure of the system. The PO clerk wears a headset microphone with a "push-to-talk" switch. The screen in front of the clerk displays a menu of topics available to him/her e.g. "Postage", "DVLA", "Bill Payments ", "Passports". Speaking any of these words invokes another screen showing a list of phrases relevant to this category which can be recognised. However, this is only an "aide-memoire" to the clerk, and all phrases are active (i.e. can be recognised) 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.

Prior to designing the system, we obtained transcripts of recordings of PO transactions at three locations in the UK, in all 16 hours of business. Inevitably, much of the dialogue transcribed was in the nature of social interaction and had little to do directly with the transaction in hand. However, analysis of these transcriptions was essential for estimating the vocabulary which would be needed by our system to achieve a reasonable coverage of the most popular transactions. At the end of this analysis, we prepared a set of 115 phrases which we estimated should be adequate to cover about 90% of transactions. After each trial, we extended the number of phrases.

2.3. **Speech recognition.** The speech recogniser used was the Entropic HAPI (HTK Application Interface) system [13], which incorporates the HTK (Hidden Markov Model

Toolkit) recogniser [7]. It is a standard HMM recogniser that first parameterises the speech signal into a sequence of vectors, each vector consisting of a set of mel-frequency cepstral coefficient (MFCC's) plus their derivatives and an energy term. The recogniser has stored speech models of several thousand "triphones" (phonemes in left and right context), each speech model consisting of a three-state hidden Markov model [4] with a multi-variate Gaussian mixture distribution of vectors associated with each state. A network of legal phrases is supplied to the recogniser, which uses a dictionary to decompose each word within a phrase into a sequence of triphones. Decoding of the speech signal is done using a Viterbi decoder that uses the speech models and the network supplied to output the most likely sequence of words given the acoustic input.

An important point about the operation of the recognition system is that both the speech models and the network can be varied. The speech models are adapted to the voice of each user using maximum-likelihood linear regression (MLLR) adaptation [11], a process which takes only a few minutes, 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 network constrains the speech recogniser 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 a number of word nodes or sub-networks, finishing with an end node (again denoting silence). Sub-networks are useful ways of defining phrase segments which can vary. For instance, a sub-network called "one2hundred" represents the legal ways of saying the integers between one and 100, and this can be inserted at any appropriate point into the network. There are other sub-networks called "amounts-of-money", "days-of-the-week", "countries" etc. By constraining the grammar in this way, recognition accuracy is significantly improved over using a looser grammar (for instance, many commercial dictation packages use a probabilistic "bigram" language model, in which the decoding of the speech utterance is constrained by the probability of any word in the vocabulary following any other word). Also, because the recogniser operates on a "best-match" basis, a phrase which is phonetically "close" but not identical with a phrase in the network will be recognised as the latter, which confers some flexibility on the speech of the clerk. (For instance "Put that on the scales, please" would be recognised as "Please put it on the scales"). The network was constructed using a graphical network-building tool, grapHvite [15]. This tool enables easy construction and editing of a network of phrases. A fragment of the network is shown in Fig 2.

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.

3. THE DEAF-SIGNING AVATAR, TESSA

The most direct way of signing phrases would be to store video-recordings of a human signer making the individual phrases, and concatenate the appropriate phrases in response to the output from the speech recogniser. However, we have been developing an experimental system that uses a virtual human (avatar) to sign teletext subtitles [19]. In this broadcast application, using an avatar has an important advantage over using video, which

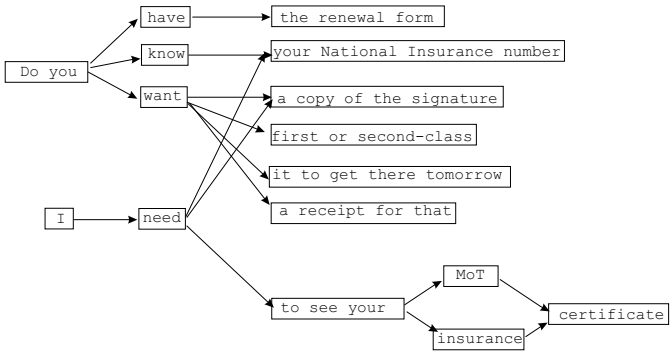

FIGURE 2. A section of the recognition network

is that the signing can be transmitted using 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 Post Office system described here, we envisage that a later, more advanced system, will use a "text-to-sign" synthesiser that will be capable of synthesising signs from unrestricted text. This clearly cannot be accomplished by concatenating video clips. Another advantage of using an avatar is that different figures can be rendered onto the avatar's frame, so that a set of recordings of signs can be used to drive different virtual humans. For these reasons, we decided to display the signs using an avatar, Tessa, which was based on the avatar developed for signing teletext subtitles.

Methods for capturing signing movements directly from video have been reported [17, 9, 2, 16] but, although this is desirable, it is an approach that is not yet practical. The alternative is to capture the signs using separate sensors for the hands, body and face, and our experience is that the fidelity and realism of movement appropriate for signing can be obtained from a virtual human using this technique.

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 space relative to a magnetic field source.
(3) Facial movements are captured using a helmet mounted camera with infra-red filters and surrounded by infra-red light emitting diodes to illuminate, typically 18, Scotchlight reflectors stuck onto the face. The 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 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

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, upper arm, body and head to track the body.

added to the lexicon then signs are manually refined, using a custom editor program, and the beginning and end of each sign marked.

The motion-data stream is displayed using a virtual human. In common with many avatars, a three-dimensional "skeleton" is driven directly from the motion-data. The "skeleton" is wrapped in, and elastically attached to, a texture mapped three-dimensional polygon mesh that is controlled by a separate thread (event loop) that tracks the "skeleton". We use one of the latest PC accelerated 3D graphics cards to render the resulting 5000 polygons at 50 frames/s using Direct-X [3] on a Pentium class PC. As a full three-dimensional model, the pose of Tessa can be changed on-the-fly by the user as can the identity of the virtual human and other characteristics. Tessa is capable of signing in real time with a refresh rate of approximately 40 frames per second. It is essential that the virtual human can be shown to convey useful information. This necessitates a range of techniques for deriving feedback and evaluating the legibility of signs. A comprehensive set of trials of the system were performed so as to evaluate its usefulness to deaf users, and these are described in section 4.

FIGURE 4. Signs for the four days, Monday, Tuesday, Wednesday, Thursday

4. EVALUATION

4.1. **Participants.** Six pre-lingually 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 the Deaf (RNID) offices. Table 4.1 gives their biographical details. They were paid for their participation and all travel and accommodation costs reimbursed.

BSL-user	Gender	Age	Age became deaf	Other methods of communication able to use
1	M	41 – 50	Birth	SSE
2	F	41 – 50	Birth	SSE
3	M	21 – 30	Birth	SSE; spoken English
4	M	21 – 30	0 - 2	SSE; spoken English
5	M	21 – 30	Birth	-
6	F	31 – 40	Birth	SSE

Three clerks were recruited by the PO to take part in the evaluations. Details of their PO working history and previous experience of communicating with deaf people in the PO are listed in Table 4.1.

Clerk	Gender	Years worked as a clerk	How often serve deaf customers	How communicate with deaf customers
1	M	11 – 20	1 a week	Speak; write things down;
2	F	11 – 20	1 a week	Speak; write things down;
3	M	11 – 20	2 to 4 a week	Speak; write things down;

4.2. **Protocol.** The evaluations took place over three sets of two days. Two BSL-users and one clerk attended for each pair of days. The first day started with completion of the first part of the questionnaire. Each BSL-user then alternated between performing a block of transactions and identifying a block of phrases. Finally, the remainder of the questionnaires were completed at the end of the second day and any general feedback recorded. BSL/English interpreters were present throughout.

4.3. **Phrase intelligibility.** The intelligibility of the phrases available in TESSA was measured in order to provide a baseline measure for assessing the signed content of future versions of the system. The deaf participants were asked to identify each phrase in the system. They were presented with each signed phrase and asked to write down what they understood. They were also asked to rate each phrase for ease of identification (on a 5-point scale from 1–"Very difficult" to 5–"Very easy") and to rate how acceptable the phrase was as an example of BSL (on a 3-point scale from 1–"Low" to 3–"High").

133 phrases were generated from the 115 distinct phrases by incorporating appropriate days of the week and numbers to ensure that each day and each number (units and tens) was presented at least once. Phrases were presented on a Gateway Solo 9300 laptop PC. The TESSA avatar operated at a rate of 25 frames per second. Signed phrases were presented without text. The BSL-user 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, amounts of money. Accuracy of identification of phrases was assessed in two ways

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

The 133 phrases gave a total of 444 sign units. While these units were not all distinct (for example, the sign for "pound" was presented several times), identification of each presentation of a unit was scored separately. One experimenter judged the accuracy of responses for both measures on the basis of written responses from each BSL-user. Once each phrase had been scored for accuracy of identification, each BSL-user 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 inappropriate or whether they were just not clear.

4.4. **Results.**

4.4.1. *Intelligibility.* 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 phrases. The average accuracy of identification of whole phrases was 61% and ranged from 42% to 70% across BSL-users (Figure 5a). For the identification of sign units in phrases, average accuracy was 80% and ranged from 67% to 89% Figure 5b) Subsequent analysis of the sign units which were wrongly identified indicated that on average 30% of errors (6% of all sign units) were due to incorrect signs and the remaining 70% (13% of all sign units) were due to unclear signing (Figure 2).

4.4.2. *Ease of identification.* Table 4.4.2 shows the percentage of phrases which were rated in each category of ease of identification. The average rating was 3.6, ranging from 2.8 to 4.5 across BSL-users. The percentage of phrases which achieved each rating from each BSL-user are shown in Figure 3.

4.4.3. *Acceptability.* Table 4.4.2 shows the percentage of phrases which were rated in each category of acceptability. The average acceptability rating was 2.2 and ranged from 1.7 to

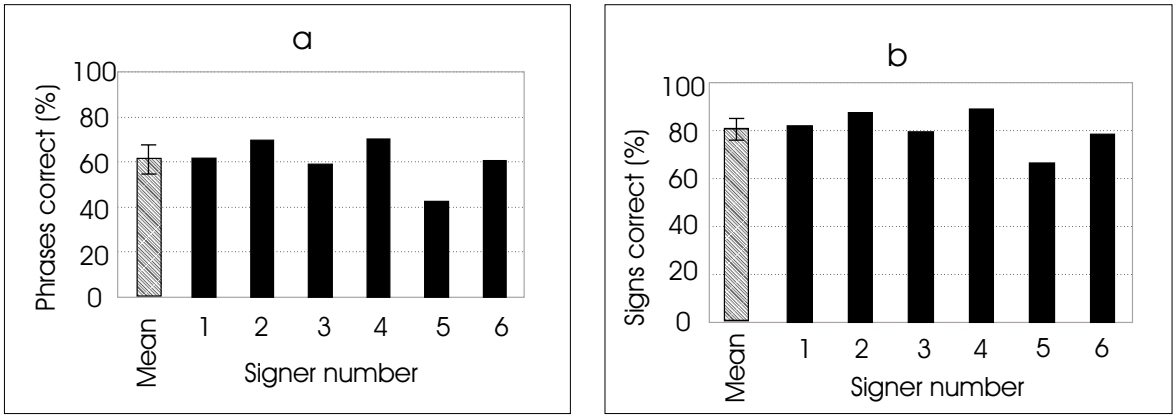

FIGURE 5. The average accuracy of identification of (a) whole phrases and (b) sign units in phrases achieved by each signer. Error bars show the 95% confidence intervals of the overall means for the 6 signers.

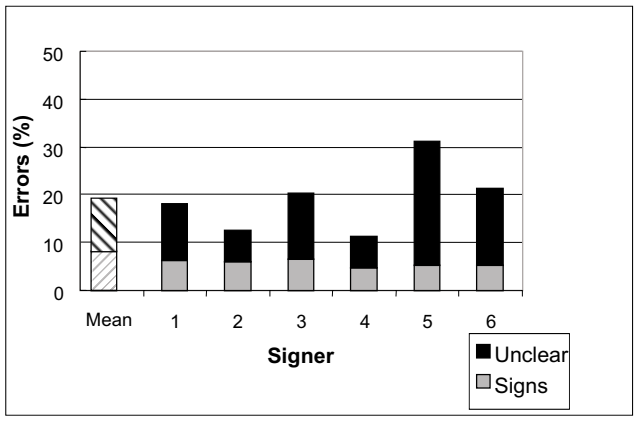

FIGURE 6. The percentage of identification errors made by each BSL-user categorised according to whether the error was due to an inappropriate sign (lower section of each bar) or was just unclear (upper section). Striped bars show the mean errors over all BSL-users.

2.8. The percentages of phrases which received each rating from each BSL-user are shown in Figure 8.

4.4.4. *Discussion.* Accuracy of identification of the signed phrases was fairly high, with averages of 61% for whole phrases and 80% for sign units, with quite a wide range in accuracy across BSL-users (ranges of 28% and 22%, 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 future, it may be more appropriate to use more than six BSL-users 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

Rating of ease of identification		% of phrases
Very easy	5	31.5
	4	26.3
	3	21.6
	2	14.0
Very difficult	1	6.7

FIGURE 7. The percentage of phrases rated in each category of ease of identification.

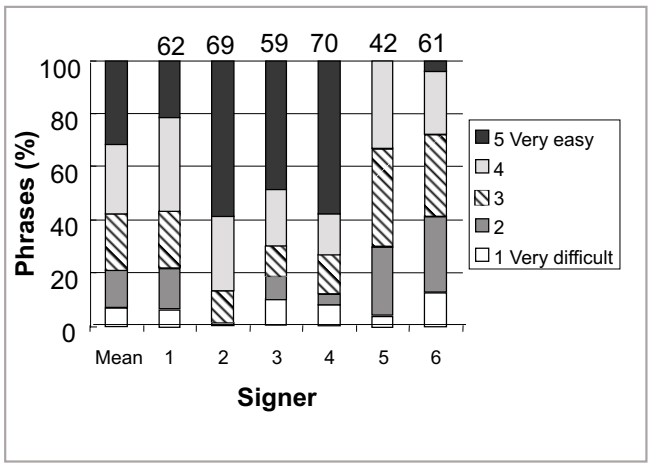

FIGURE 8. Percentage of ratings made by each BSL-user for ease of identification of phrases in each category on a 5-point scale ranging from 1- "Very difficult" to 5- "Very easy". Percentage values above each bar show the accuracy of identification of whole phrases achieved by each BSL-user.

greatly between subjects, with personal averages ranging from 4.9% to 6.6%. This pattern might suggest that the same signs were considered inappropriate by all BSL-users. However, inspection of the pattern of errors across BSL-users for each phrase indicated that this was not necessarily the case. Of the 46 phrases where one or more sign was considered inappropriate by any BSL-user, in 34 of these (74%) a sign was considered inappropriate by no more than two of the BSL-users. This result suggests that regional variations or differences in personal signing style may have played a role in phrase intelligibility. While over half the phrases were considered easy to identify in the top two categories of the 5-point rating scale, and 77% in the top three categories, each BSL-user rated phrases across at least 3 categories (Figure 3). This range of rating indicates that all BSL-users perceived differences between the phrases in terms of ease of identification. 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. Hence there is scope for improvement in both the subjective ease of identification and acceptability of phrases as exemplars of BSL.

As might be expected, there appears to be some relationship between average identification accuracy and ratings of ease of identification and acceptability for each BSL-user (Figure 3). For example, BSL-users 2 and 4 who achieved the highest levels of accuracy

rated more phrases as "Very easy" than BSL-user 5 who achieved the lowest level of accuracy. However, accuracy does not account for all variation in ratings. These two subjective measures therefore appear to provide additional information to numerical measures of intelligibility.

4.4.5. *Transactions.* Staged transactions were used to compare completion times and ease and acceptability of communication with and without TESSA. Each BSL-user attempted 18 transactions with a single PO clerk. Transactions were selected by the PO as those achievable with the phrases available. Of the 18 transactions, 6 were denoted simple, 6 average difficulty and 6 complex. The average difficulty and complex transactions were attempted twice by each BSL-user/clerk pair, once with an open counter and once behind a "fortress" 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.

The actual complexity of the transactions varied between clerks according to the number of questions they asked and assumptions they made. Hence differences between performances for each designated category of complexity were not considered valid for analysis post hoc. Half of all transactions were attempted with TESSA and half without. The phrases presented with/without TESSA were counter-balanced between BSL-users. Three practice transactions were performed with TESSA at the start of each session so that the clerk, BSL-user and interpreter could get used to using TESSA and the format of the evaluation. Transactions were performed in blocks of 6, three with TESSA and three without.

Transactions were performed using a Gateway Solo 9300 laptop PC enabling a presentation rate of 25 frames per second. The clerks used a hand-held Apple microphone. The speech recognition system was trained by each clerk for half-an-hour on the first day of testing before using the system in the trials. Sound levels were calibrated at the start of each day of testing. The approximate time taken to successfully complete each transaction was recorded. On completion of each transaction, both BSL-users and clerks were asked to rate each transaction for ease of communication (on a 5-point scale from 1–"Very difficult" to 5–"Very easy") and acceptability (on a 3-point scale from 1–"Low" to 3–"High").

On average, transactions took longer to complete with TESSA than without [F(1,178)=61.2, p¡0.001] (Figure 5). Average times for transactions were 57s without TESSA and 112s with TESSA. On average, transactions completed with TESSA were rated as more difficult than transactions completed without TESSA [U(1,178)=6009, p¡0.001] (Figure 6). On the 5-point scale (from 1–"Very difficult" to 5–"Very easy") average ratings were 4 with TESSA and 4.3 without. On average, communication in transactions completed with TESSA was rated as less acceptable than in transactions completed without TESSA [U(1,178)=6025, p¡0.001] (Figure 7). On the 3-point scale (from 1–"Low" to 3–"High") average ratings were 1.9 with TESSA and 2.6 without. Clerks rated communication in transactions completed with TESSA as slightly more difficult than transactions completed without Tessa [U(1,178)=5232, p¡0.001] (Figure 8). On the 5-point scale (from 1–"Very difficult" to 5–"Very easy") average ratings were 4.0 with Tessa and 4.4 without. They also rated communication in transactions completed with Tessa as less acceptable than in transactions completed without Tessa [F(1,178)=89, p¡0.05] (Figure 9). On a 3-point scale (from 1 -"Low" to 3 -"High") average ratings were 2.5 with Tessa and 2.6 without.

4.4.6. *Discussion.* Compared to transactions without TESSA, transactions performed with TESSA on average took twice as long to complete and the deaf participants, and to a lesser extent the clerks, rated communication as more difficult and less acceptable. The main

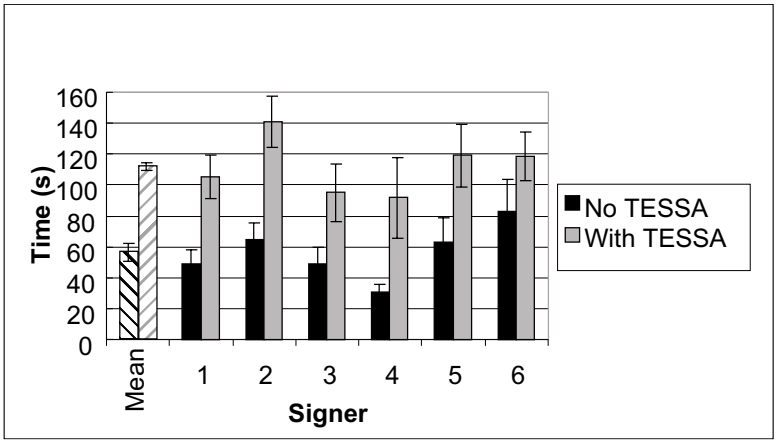

FIGURE 9. Average times taken for transactions without TESSA (light-coloured bars) and with TESSA (dark bars), for each BSL-user and over all BSL-users. Error bars show the 95% confidence intervals of the means. BSL-user rating for ease of communication.

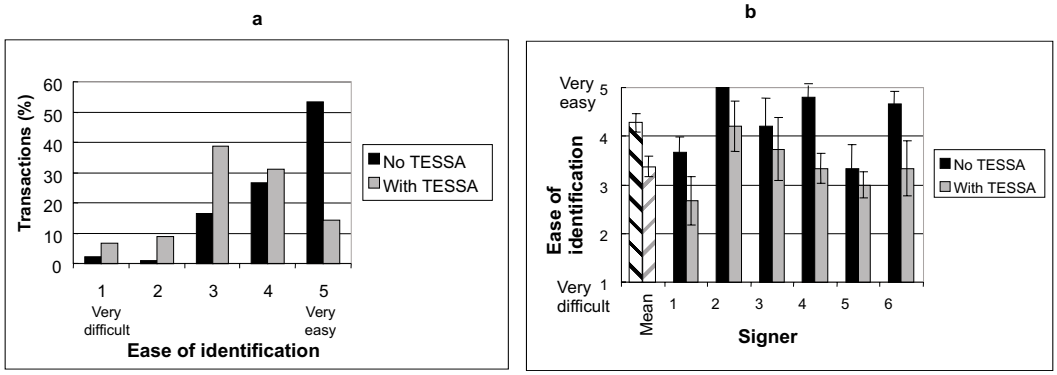

FIGURE 10. Ratings of ease of communication in each category on a 5-point scale from 1-"Very difficult" to 5- "Very easy" in transactions without TESSA (light bars) and with TESSA (dark bars). (a) Percentage of transactions rated in each category. (b) Average ratings for each BSL-user and over all BSL-users. Error bars show the 95% confidence intervals of the means.

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 on a list 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 are likely to decrease substantially with training and experience on the system, and moreover with use of the next, unconstrained version of the system where phrases do not need to be repeated verbatim.

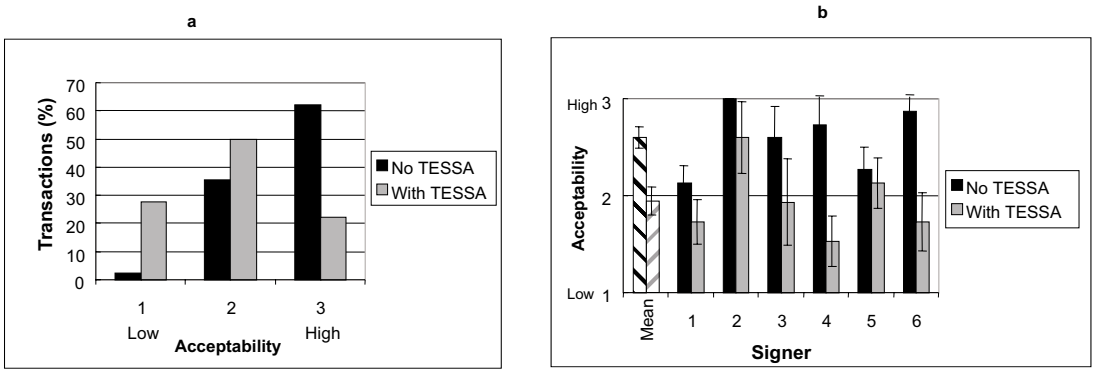

FIGURE 11. Ratings by BSL-users for acceptability on a 3-point scale from 1-"Low" to 3- "High" of transactions without TESSA (light bars) and with TESSA (dark bars). (a) Percentage of transactions rated in each category. (b) Average ratings for each BSL-user and over all BSL-users. Error bars show the 95% confidence intervals of the means.

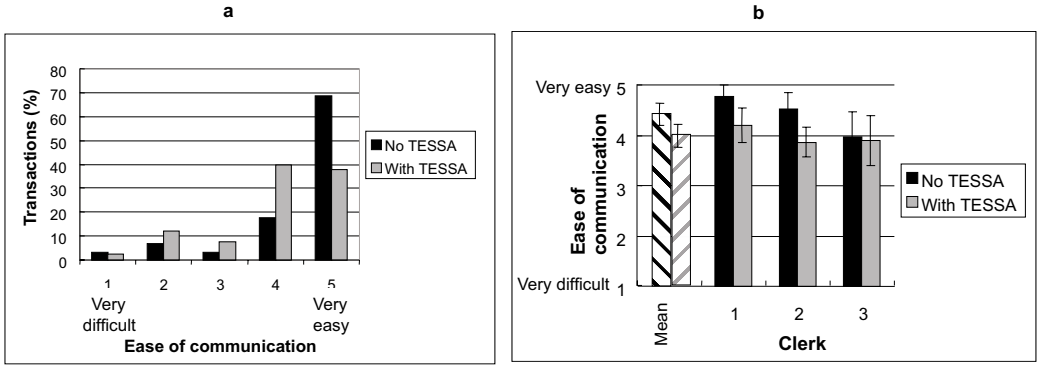

FIGURE 12. Rating of communication by clerks on a 5-point scale from 1-"Very difficult" to 5- "Very easy" for transactions without TESSA (light bars) and with TESSA (dark bars). (a) Percentage of transactions rated in each category. (b) Average ratings from each clerk and the overall mean rating. Error bars show the 95% confidence intervals of the means.

Additional factors may have contributed to the longer transaction times and poorer ratings with TESSA. First, the list of phrases were 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 giro cheque or claiming a pension payment. As transactions used in this evaluation were limited by the phrases available, they also tended to be fairly simple or were simplified. This was confirmed by the PO staff who selected the transactions and the clerks who often 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

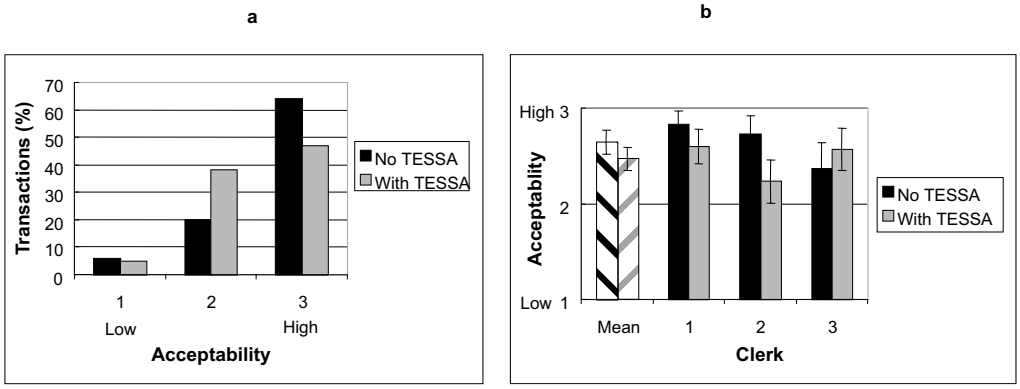

FIGURE 13. Ratings by clerks for acceptability on a 3-point scale from 1-"Low" to 3- "High" of transactions without TESSA (light bars) and with TESSA (dark bars). (a) Percentage of transactions rated in each category. (b) Average ratings for each clerk and over all clerks. Error bars show the 95% confidence intervals of the means.

was fairly easy without TESSA. Second, 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 where necessary. This is a consequence of the type of people who would be prepared to attend two days of testing away from home, 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. Third, the clerks either were deaf aware or soon became deaf aware as a result of spending two 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. Fourth, 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.

4.4.7. *Questionnaires.* Questionnaires were used to obtain subjective views of previous experiences of communication in the PO. In addition, both BSL users and PO clerks were asked for general views on the use of avatars for sign-language communication. The BSL-users' questionnaire included use of three visual analogue scales where BSL-users were asked to rate clarity of signing, acceptability of the avatar as a signer of BSL and the avatar appearance. This was achieved by marking on a line with "Not at all" (0%) at one end and "Totally" (100%) at the other end. Location of marked points on the line yielded a score between 0% and 100% on each scale for each BSL-user.

All BSL-users said that if they needed something from the PO they would usually go themselves and would attempt to communicate by lip-reading/speaking and gesturing. Five BSL-users said they would also write things down and ask the clerk to write things down if needed. BSL-user 5 said he would take someone to interpret if communication became difficult and usually refused to use written communication (although he did use written communication in some of the trials without TESSA in the evaluation). Results from the

three questions asking about ease of communication in the PO, previously, in the trials with TESSA and anticipated in everyday life with TESSA are shown for each of the six BSL-users in Table 5.

For the trials, ratings stayed the same or were worse for all but BSL-user 1 who said communication was "Very difficult" previously but "Manageable" with TESSA in the trials and anticipated to be "Fairly easy" with TESSA for everyday life. When asked about the relative ease of communication with and without TESSA, in the trials and anticipated in everyday life (Table 6), four and five BSL-users (respectively) said it would be "Slightly worse" or "Much worse" with TESSA. BSL-user 1, who said that communication in the PO was usually "Very difficult", said it would be "Slightly easier" with TESSA (for both questions). When asked about how much communication in the PO usually bothered them, three of the BSL-users said "A little" or "Not at all" (Table 7). When asked about communication with TESSA in everyday life, five of the BSL-users said this would bother them "A little" or "Not at all". The other BSL-user was not bothered at all previously but anticipated that using TESSA in everyday life would bother him "Quite a lot".

Question	**Very easy**	**Fairly easy**	**Manage-able**	**Slightly difficult**	**Very difficult**
How easy do you usually find communication in the Post Office?	2		3 4 6		1 5
How easy did you find communication using TESSA?			1 2 3 4	6	5
In everyday life, how easy do you think communication would be using TESSA?		1 2	6	4	3 5

Table 5. Responses made by each BSL-user to the three questions about ease of communication in the PO: previously, in the trials with TESSA and anticipated in everyday life with TESSA. Each number represents the responses from one BSL-user.

Question	**Much easier**	**Slightly easier**	**No difference**	**Slightly worse**	**Much worse**
Compared to communication without, do you think TESSA made communication:		1	3	2 4	5 6
In everyday life, do you think that using TESSA in the Post Office would make communication:		1	2	3 4 5	6

Table 6. Responses made by each BSL-user to the two questions about comparing communication in the PO with and without TESSA: in the trials and anticipated in everyday life. Each number represents the responses from one BSL-user.

Question	**Very much**	**Quite a lot**	**Some**	**A little**	**Not at all**
In everyday life, how much does communication in the Post Office upset, annoy or worry you?	1	4	3	6	2 5
In everyday life, how much would communication using TESSA in the Post Office upset, annoy or worry you?		5		6	1 2 3 4

Table 7. Responses made by each BSL-user to the two questions about how much communication in the PO bothered them, previously and anticipated with TESSA in everyday life. Each number represents the responses from one BSL-user.

When asked for a preference, four BSL-users said they would prefer to communicate without TESSA and two preferred with, as an option if needed. All BSL-users said they would prefer TESSA with both BSL and text, rather than just BSL or just text alone. Using visual analogue scales (Figure 10), the mean rating for clarity of signing was 28% and ranged from 12% to 46% across BSL-users. For acceptability of the avatar as a signer of BSL, the mean rating was 32%, ranging from 3% to 65%. The mean rating for appearance of the avatar was 30%, ranging from 2% to 66%. All clerks rated usual communication

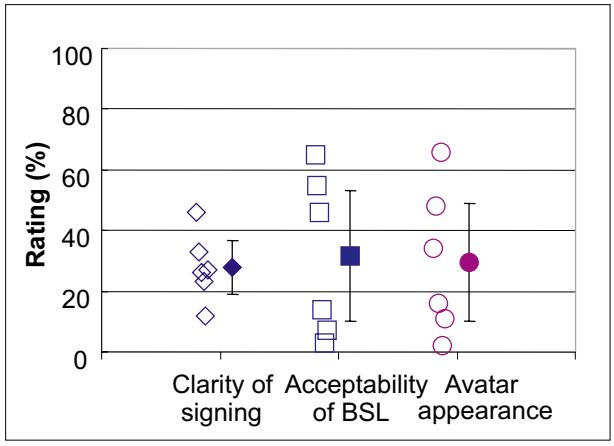

FIGURE 14. Ratings on visual analogue scales for clarity of signing (Qu.19), acceptability of avatar as a signer of BSL (Qu.20) and avatar appearance (Qu.21). Open symbols show ratings for each BSL-user with filled symbols showing the average rating. Error bars show the 95% confidence intervals of the means.

with deaf customers as "Fairly easy" and communication with TESSA as "Fairly easy" or "Very easy" (Table 8). All clerks said communication was "Slightly easier" or "Much easier" with TESSA than without, and that in everyday life they anticipated that communication would be "Much easier" with TESSA (Table 9). 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.

Question	**Very easy**	**Fairly easy**	**Manage-able**	**Slightly difficult**	**Very difficult**
How easy do you usually find communication with deaf customers?		1 2 3			
How easy did you find communication using TESSA?	3	1 2			
In everyday life, how easy do you think communication would be using TESSA?	2 3	1			

Table 8. Responses made by each clerk to the three questions about ease of communication with deaf customers: previously, in the trials with TESSA and anticipated in everyday life with TESSA. Each number represents the responses from one clerk.

Question	**Much easier**	**Slightly easier**	**No difference**	**Slightly worse**	**Much worse**
Compared to communication without, do you think TESSA made communication:	3	1 2			
In everyday life, do you think that using TESSA in the Post Office would make communication:	1 2 3				

Table 9. Responses made by each BSL-user to the two questions about comparing communication in the Post Office with and without TESSA, in the trials and anticipated in everyday life. Each number represents the responses from one clerk.

5. Discussion of evaluation results

Of the six deaf participants, one person said that communication would be easier with TESSA and two BSL-users said they would prefer to communicate with TESSA in the PO, as an option in case communication became difficult. The three deaf participants who said that communication in the PO usually upset or worried them, said they thought using TESSA in the PO would not bother them at all. While this represents positive feedback from some BSL-users, the fact that these responses were not more generally positive does not seem unreasonable at this stage in the life-cycle of the project. These 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, where communication with TESSA was somewhat lengthy and disjointed.

Scores on the visual analogue scales showed a wide range of responses between BSL-users for ratings of clarity of signing, acceptability of the avatar as a signer of BSL and the avatar's appearance. These scales proved easy to use as the BSL-users responded more easily than to the previous questions for which it was often difficult to obtain a categorical response. These scales therefore are likely to be useful outcome measures for evaluating future signing avatars and versions of the system. Scores on the scales were all under about 65, hence there is much scope for improvement.

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

- Facial expressions need to be improved.
- Clearer handshapes, finger configurations and lip patterns are required, especially for numbers and finger-spelling.
- The delay between the end of the spoken phrase and the beginning of signing needs to be reduced.
- The appearance of the avatar needs to improve. In particular, a clearer distinction between should be made between the face and hands and the clothing, which should be plain.
- All BSL-users 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.

Comments on use of avatars for signing in general included the following:

- The deaf participants found avatars more useful for more complex communication needs, e.g. explaining benefits forms.
- 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 would ensure more frequent use and hence greater likelihood that the system would be used with deaf people.
- Clerks also commented that they would like more phrases and an unconstrained speech system, where phrases need not be spoken verbatim.

5.1. **General comments and future work.** Our goal in developing this trial system was to establish whether the introduction of a limited speech-to-sign translation system for the Post Office counter clerk would be beneficial to deaf users who used signing as their primary form of communication. Although some of the feedback from the evaluation was critical, we are encouraged by the following points:

- Two of the six deaf participants said they would prefer to have TESSA available in the PO for use when communication became difficult. The other four said they would prefer to communicate without TESSA in its present form.
- One of the deaf participants said that communication in the PO would be easier with TESSA..
- The three deaf participants who said that communication in the PO usually upset or worried them said they thought using TESSA in the PO would not bother them at all.
- Feedback from the PO clerks was generally very positive, despite the very limited time they had to train with the system.

The evaluations have indicated that there is much scope for improvement of TESSA, gave some insight into how these improvements could be achieved and provided baseline outcome measures against which improvements could be assessed. The majority of aspects identified for improvement are planned for further development within the ViSiCAST project. Primarily, the development of an unconstrained version, where phrases need not be repeated word for word, would enable much more natural communication and should greatly reduce the time taken for transactions, so is also likely to be more acceptable to both deaf customers and clerks. Other aspects to be explored include research into facial modelling which should improve avatar facial expressions and lip patterns. New data gloves are also being used to improve recording of finger movements and handshapes. New models of the avatar and clothing will also take account of the comments made by the deaf participants. Less formal evaluations are planned within the deaf community to assess the views of more deaf people, but further formal evaluations will continue through the lifetime of the ViSiCAST project.

REFERENCES

[1] J.A. Bangham, S.J. Cox, M. Lincoln, I. Marshall, M. Tutt, and M. Wells. Signing for the deaf using virtual humans. In *IEE Colloquium on Speech and language processing for disabled and elderly people*, April 2000.

[2] Lien C C and Huang C L. Model-based articulated hand motion tracking for gesture recognition. *IMAGE AND VISION COMPUTING*, 16(2):121 – 134, Feb 1998.

[3] Microsoft Corporation. Directx home page. http://www.microsoft.com/directx.

[4] S.J. Cox. Hidden Markov Models for automatic speech recognition: theory and application. In C Wheddon and R Linggard, editors, *Speech and Language Processing*, pages 209–230. Chapman and Hall, 1990.

[5] Brien D. *Dictionary of British Sign Language / English*. Faber and Faber, 1992.

[6] Klima E. and Bellugi U. *The Signs of Language*. Harvard University Press, 1979.

[7] J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK book*. Entropic Research Laboratories Inc., 1996.
[8] R.D. Johnston et al. Current and experimental applications of speech technology for Telecom services in Europe. *Speech Communication*, 23(1–2):5–16, 1997.
[9] Huang C L and Huang W Y. Sign language recognition using model-based tracking and a 3d hopfield neural network. *Machine vision and applications*, 10(5-6):292 – 307, Apr 1998.
[10] L.F. Lamel et al. The LIMSI RailTel system: field trial of a telephone service for rail travel information. *Speech Communication*, 23(1–2):67–82, 1997.
[11] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9(2):171–85, 1995.
[12] B. Mazor and B. L. Zeigler. The design of speech-interactive dialogs for transaction- automation systems. *Speech Communication*, 17:313–320, November 1995.
[13] J. Odell, D. Ollason, V. Valtchev, and D. Whitehouse. *The HAPI Book*. Entropic Cambridge Research Laboratory, 1997.
[14] F. Pezeshkpour, I. Marshall, R. Elliott, and J.A. Bangham. Developing of a legible deaf signing virtual human. In *IEEE Multimedia Systems Conference '99 (IEEE ICMCS'99)*, pages 333–338, June 1999.
[15] K. Power, R. Morton, C. Matheson, and D. Ollason. *The grapHvite Book*. Entropic Cambridge Research Laboratory, 1997.
[16] Ahmad T, Taylor CJ, Lanitis A, and Cootes TF. Tracking and recognising hand gestures, using statistical shape models. *IMAGE AND VISION COMPUTING*, 15(5):345 – 352, May 1997.
[17] Starner T, Weaver J, and Pentland A. Real-time american sign language recognition using desk and wearable computer based video. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 20(12):1371 – 1375, Dec 1998.
[18] A. Waibel. Interactive translation of conversational speech. *Computer*, 29(7), 1996.
[19] M. Wells, F. Pezeshkpour, M. Tutt, J.A. Bangham, and I. Marshall. Simon - an innovative approach to deaf signing on television. In *Proc. International Broadcasting Convention*, pages 477–482, September 1999.
[20] P.J Wyard et al. Spoken language systems—beyond prompt and response. In F.A. Westall, R.D. Johnston, and A.V. Lewis, editors, *Speech Technology for Communications*, pages 487–520. Chapman and Hall, 1998.
[21] O. Yoshioka, Y. Minami, and K. Shikano. A speech dialogue system with multimodal interface for telephone directory assistance. *IEICE ransactions on Information and Systems*, E78D(6):616–621, 1995.

[1]SCHOOL OF INFORMATION SYSTEMS, UNIVERSITY OF EAST ANGLIA, NORWICH NR4 7TJ, U.K,
[2]ROYAL NATIONAL INSTITUTE FOR DEAF PEOPLE, 19–23 FEATHERSTONE STREET, LONDON EC1Y 8SL, U.K.

E-mail address: {ml,sjc}@sys.uea.ac.uk, melanie.nakisa@rnid.org.uk