PRIMERA EXPERIENCIA DE TRADUCCIÓN AUTOMÁTICA DE VOZ EN LENGUA DE SIGNOS

A. Huerta, E. Ibáñez, R. San-Segundo, F. Fernández, R. Barra, L.F. D'Haro
Speech Technology Group. Universidad Politécnica de Madrid. Spain

Resumen

Este artículo presenta los primeros experimentos de traducción de voz a Lengua de Signos Española (LSE). El sistema desarrollado está centrado en un dominio concreto de aplicación: frases pronunciadas por un funcionario que atiende a las personas que desean obtener o renovar el DNI (Documento Nacional de Identidad) o el pasaporte. El sistema reconoce y traduce las frases pronunciadas por el funcionario a lengua de signos. El sistema está formado por 3 módulos principales: un reconocedor de voz, un módulo de traducción de lenguaje natural (traduce secuencias de palabras en secuencia de signos), y finalmente un módulo de representación de los signos mediante un agente animado en 3D. Los mejores resultados obtenidos ofrecen un error en la secuencia de signos del 24.0% y un BLEU (BiLingual Evaluation Understudy) del 0.70.

1. Introducción

La lengua de signos presenta una gran variabilidad entre diferentes países e incluso en diferentes áreas de un mismo país. Por esta razón, desde 1960, la lengua de signos ha sido objeto de estudio en Estados Unidos [1][2][3], en Europa [4][5], Africa [6] y Japón.

En España, durante los últimos 20 años ha habido varias propuestas para normalizar la Lengua de Signos Española (LSE), pero ninguna de ellas ha sido aceptada por las personas sordas. Desde su punto de vista, estas propuestas tienden a limitar la flexibilidad de la lengua de signos. En 1991 M.A. Rodríguez [8] llevó acabo un análisis detallado de las diferencias entre la lengua de signos utilizada por las personas sordomudas y las propuestas de estandarización. Este estudio constituye uno de los principales estudios de la LSE y ha sido una referencia importante en este trabajo.

La traducción de habla natural ha sido y sigue siéndolo hoy en día una de las principales áreas de investigación en proyectos como C-Star, ATR. Vermobil, Eutrans, LC-Star, PF-Star y TC-Star. Todos estos proyectos, a excepción de TC-Start, son proyectos enfocados a la traducción de habla en dominios restringidos (como por ejemplo una agencia de viajes o turismo) con vocabularios de tamaño medio (10.000 palabras). Los mejores sistemas de traducción están basados en soluciones estadísticas [9], incluyendo técnicas basadas en ejemplos [10], traductores de estados finitos [11] y otras soluciones basadas en datos. Los importantes progresos conseguidos en traducción de habla se deben principalmente a factores como la aparición de medidas de error [12], mejora de la eficiencia de los algoritmos de entrenamiento [13], desarrollo de modelos dependientes del contexto [10] y algoritmos de generación eficientes [14].

El proyecto europeo eSIGN (Essential Sign Language Information on Government Networks) [15] constituye uno de los esfuerzos más importantes para el desarrollo de herramientas de apoyo a la generación de contenidos para personas sordas. En este proyecto, uno de los principales resultados ha sido la generación de un avatar en 3D (VGuido) (capaz de representar cualquier signo de la lengua de signos) y un entorno visual para el desarrollo de animaciones fácilmente. Estas herramientas están orientadas a traducir contenido web. La lengua de signos es la lengua principal para muchas personas sordas (algunas de las cuales no comprenden la lengua escrita puesto que son personas sordas prelocutivas). El resultado del proyecto eSIGN está funcionando en páginas web oficiales de Alemania, Países Bajos y el Reino Unido.

En los últimos años ha aumentado el interés de varios grupos por la traducción automática a lengua de signos desarrollándose varios prototipos: basados en ejemplos [16], basados en reglas [17], frases completas [18] o soluciones estadísticas [19]. Este artículo presenta unos primeros experimentos sobre uno de los primeros sistemas de traducción de voz a signos y el primero para Lengua de Signos Española.

2. Descripción del sistema

El sistema está formado por varias partes como se puede ver en la Figura 1.

2.1. Un reconocedor de voz.

Se encarga de traducir la voz en una secuencia de palabras. Este módulo permite reconocer habla en lenguaje natural (habla continua) e independiente del locutor. Para esta aplicación el vocabulario posible contiene 485 palabras, lo que permite conseguir tasas de aciertos por encima del 90%.

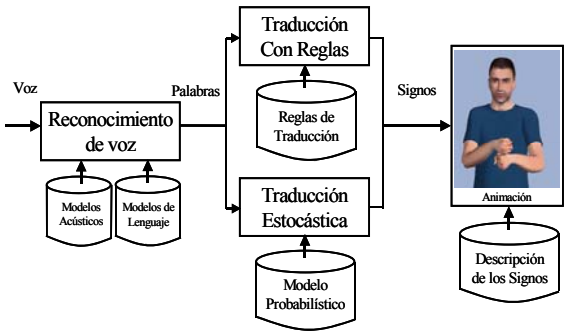

Figura 1. Diagrama de módulos del sistema desarrollado

2.2. Un módulo de traducción automática.
Traduce la secuencia de palabras en una secuencia de signos de la LSE. Se han utilizado dos alternativas tecnológicas para la traducción. La primera de ellas se basa en reglas de traducción elaboradas por una persona experta. Mediante este método, la traducción entre palabras y signos se realiza de forma manual: las reglas van combinando palabras y generando signos, que a su vez se pueden recombinar con otros signos o palabras para formar la secuencia completa final.

La segunda alternativa para la traducción se basa en métodos estadísticos cuyos modelos se aprenden a partir de un corpus paralelo de frases texto-secuencias de signos. Para el desarrollo del sistema se dispone de un conjunto de 201 frases en texto y su traducción en LSE.

2.3. Un módulo de representación de los signos.

Este módulo está basado en un agente animado en 3D. Se dispone de dos agentes animados, que podrán usarse de forma indistinta. Uno de ellos es 'VGuido', desarrollado en el proyecto eSIGN, y el segundo 'Sara', que ha sido desarrollado en este trabajo. 'Sara' ha sido desarrollada atendiendo a las necesidades de perfeccionamiento de las partes del cuerpo determinantes a la hora de representar un signo de la lengua de signos (cabeza y brazos). Asimismo, se ha implementado una interfaz gráfica capaz de general signos, posiciones y animaciones, que funciona como módulo de representación del sistema desarrollado, y también como sistema independiente de diseño de animaciones para cada signo, al servicio de un logopeda, con la posibilidad de incrementar el número de signos a realizar, creando bibliotecas específicas. Para ello se ha dotado al sistema gráfico de herramientas y funcionalidades de alto nivel que permiten desarrollar nuevas animaciones con cierta facilidad para el usuario.

3. Dominio de aplicación

Los módulos de análisis semántico y de generación de signos de este sistema están diseñados para el dominio de aplicación de apoyo al servicio de solicitud/ renovación del DNI (Documento Nacional de Identidad) o pasaporte. Dicho proceso se gestiona desde el Ministerio del Interior, y por lo tanto, toda la información que se ha usado para el análisis de este proceso de solicitud y renovación del DNI se ha obtenido de la página oficial del Ministerio de Interior.

A partir de la información de dicha página se redactaron 201 frases. Se ha mantenido por lo general la estructura de las frases originales, poniendo mucho interés en generar diferentes niveles de uso, o grados de flexibilidad a la hora de realizar la traducción del español a la LSE. Si bien es cierto que, por lo general, se ha tratado de simular los hábitos de las personas sordomudas que prefieren dividir una misma oración en una serie de sentencias mucho más cortas y directas. Esto ha supuesto una reducción sustancial del tamaño de algunas de las oraciones compuestas, tal y como se muestra en el siguiente ejemplo:
Frase original: "El documento tramitado lo recogerá su titular o persona que autorice mediante la presentación del correspondiente resguardo, en la misma Oficina donde lo solicitó."
Algunas de las frases generadas a partir de la original:

- "Para recoger el denei tienes que ser el titular o tener su permiso."
- "Para recoger el denei deberá presentar el correspondiente resguardo."
- "Para recoger el denei debes ir allí donde lo solicitaste."

El sistema pretende, por lo tanto, conseguir que un funcionario de la administración del Estado que no conozca en absoluto la LSE pueda explicar a una persona sorda qué necesita para poder solicitar o renovar su DNI, cuales son las utilidades de dicho documento, dónde puede viajar con él, y todo tipo de información relativa al DNI o el pasaporte.

4. Reconocedor de voz

Para la elaboración de los módulos de reconocimiento de voz y de traducción se ha utilizado la plataforma para el desarrollo de aplicaciones con voz del Grupo de Tecnología del Habla (GTH) denominada SERVIVOX [20].

El módulo de reconocimiento de voz convierte las frases pronunciadas en lenguaje natural a texto. Para la implementación de dicho módulo se ha usado un reconocedor de habla continua, desarrollado en el GTH. Dicho reconocedor utiliza modelos de Markov continuos, modelo de lenguaje 2-gram, 16 modelos de ruido para hacer frente a los ruidos que puedan producirse (ruidos ambiente, ruidos del locutor como carraspeos, muletillas, etc.), y un diccionario de 458 palabras. La gramática necesaria para dicho reconocedor ha sido obtenida a partir de una lista con 201 frases del dominio de aplicación. Para crear dicha gramática se usó un programa de creación de gramáticas desarrollado en el GTH.
Para comprobar el efecto del Reconocimiento Automático del Habla (RAH) en el sistema se

procedió a la lectura de las frases del dominio de aplicación para su reconocimiento automático. Así se obtuvieron por escrito tanto las transcripciones como las frases reconocidas.

5. Traducción por reglas

La primera estrategia de traducción considerada se basa en reglas de traducción desarrolladas por una persona experta. En este caso, las relaciones entre los signos y las palabras se definen manualmente. Las reglas van combinando palabras y generando signos, que a su vez se pueden recombinar con otros signos o palabras para dar lugar a la secuencia de signos final. El sistema de traducción por reglas implementado podría clasificarse como un híbrido entre la traducción directa (posee un tipo de reglas que traducen una palabra por su signo equivalente en LSE), y la traducción indirecta con transferencia de tipo sintáctico (se ha creado un modelo sintáctico del español y otro de la LSE para hacer corresponder uno con otro).

El primer paso para la realización de este sistema de traducción ha sido un análisis pormenorizado de la LSE, y la correspondiente extracción e implementación de las reglas sintácticas. En este proceso fue fundamental la referencia [8].

Durante el segundo paso se lleva a cabo la traducción según las reglas sintácticas anteriormente elaboradas. Dicho proceso se ha descompuesto en las siguientes etapas:

5.1. Categorización.

En primer lugar se asigna una o varias categorías sintáctico-semánticas a cada una de las palabras que forman el vocabulario. La codificación utilizada para dichas categorías contiene una serie de prefijos y sufijos, tal y como se muestra en la Figura 2.

Lo fundamental de este sistema de categorías es la capacidad que proporciona para discernir si la palabra es un verbo (v2s, v3s, v3p, ver) o un sustantivo (por, s2s, s3s, s3p), si puede ser un sujeto (s2s, s3s, s3p) o no (por), información del tiempo verbal (presente por defecto, pasado o futuro), del número (singular por defecto, plural), si es una palabra de traducción directa (directo) o si por el contrario debe sufrir un proceso de deletreo (deletreo) como es el caso de los nombres propios o las siglas. El resto de palabras (que no se clasifican en los grupos anteriores) se codifican con el prefijo pal y sin sufijo.

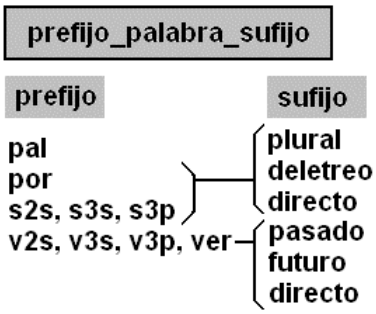

Figura 2. Sistema de categorías

5.2. Implementación de las reglas antes de eliminar basura.

Antes de eliminar las palabras basura (palabras que no aportan información en el proceso de traducción) se aplica un conjunto importante de las reglas desarrolladas. Estas reglas se pueden clasificar en 4 tipos que se describen a continuación:

- Una palabra genera un signo las palabras que se traducen según esta regla están categorizadas como 'por_palabra_directo', y su traducción es PALABRA (misma palabra en mayúsculas).
- Varias palabras generan un signo para este tipo de reglas se han usado las primitivas 'reescribeX', donde X puede ser 2, 3, 4, 5, 6 y 7, cuyo cometido es la reagrupación de los X bloques o tokens (que pueden ser palabras, signos o signos intermedios) iniciales en un único bloque o token final.
- Una palabra da lugar a varios signos para estas reglas se han usado las primitivas 'desdoblaX' (donde X puede ser 2, 3, 4, 5 ó 6), desdobla2cond (que permite extraer información del siguiente bloque), traduce_deletreos (que obtiene a partir de una palabra los signos necesarios para su deletreo), suj_omitidos (que añade un bloque para el sujeto de la frase, imprescindible en LSE), y plur_pas_fut (que añade los signos necesarios para indicar que la palabra siguiente está en plural, o que el verbo tiene información de pasado o futuro). Este tipo de reglas pretende dividir un bloque o token inicial en tantos como sean necesarios según la regla de traducción aplicada.
- Varias palabras se traducen en varios signos para esta regla se implementaron las primitivas 'reescribeXaY' (cuando el número de bloques finales es distinto del inicial) y 'reescribeXaX' (en el caso de que se obtengan tantos bloques como había inicialmente). Este tipo de reglas surge de la necesidad de resolver las situaciones en las que una expresión en el lenguaje natural genera una expresión gestual (y no un único signo).

5.3. Eliminación de las palabras basura.

Eliminación de las palabras basura o con prefijo pal_. Se eliminan los bloques que no aportan información a la frase a traducir: a bien fueron categorizados como basura inicialmente o nunca han formado parte de una regla anterior y por tanto no son útiles para la traducción.

5.4. Ordenación sintáctica de los signos/gestos y generación de la secuencia final.

Esta ordenación se lleva a cabo gracias a las primitivas plur_pas_fut (que crea los signos indicativos de plural, pasado y futuro y coloca delante del sustantivo el signo indicativo de plural), pon_primero (que trata de colocar al comienzo de la frase los signos indicativos de PASADO y FUTURO) y posesivos (que coloca los signos que signan el posesivo detrás del signo que les sigue). El proceso completo de traducción queda reflejado en la Figura 3.

6. Traducción estadística

La traducción estadística consiste en utilizar un modelo probabilístico para obtener la mejor secuencia de signos resultado de la traducción de una secuencia de palabras obtenidas del reconocedor de voz. Este modelo esta formado por dos tipos de probabilidades:

- Probabilidad de traducción. Recoge la información sobre qué palabras se traducen por qué signos.
- Probabilidad de la secuencia de signos. Aporta información sobre qué secuencias de signos son más probables en la Lengua de Signos Española.

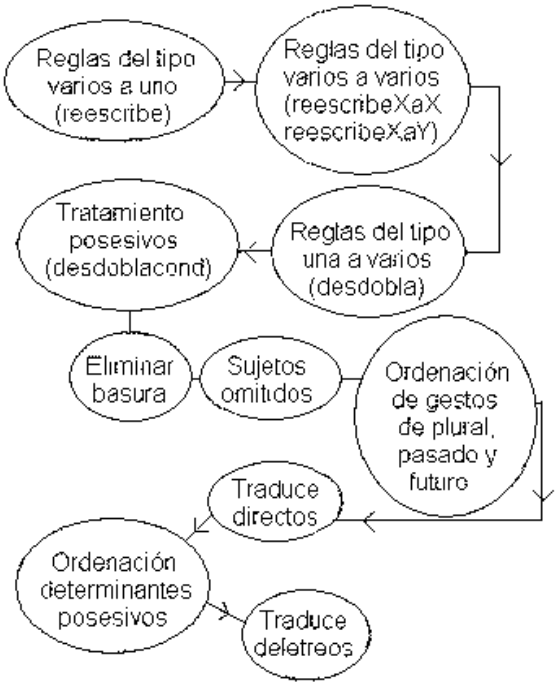

Figura 3. Esquema de las reglas y primitivas utilizadas en el proceso de traducción.

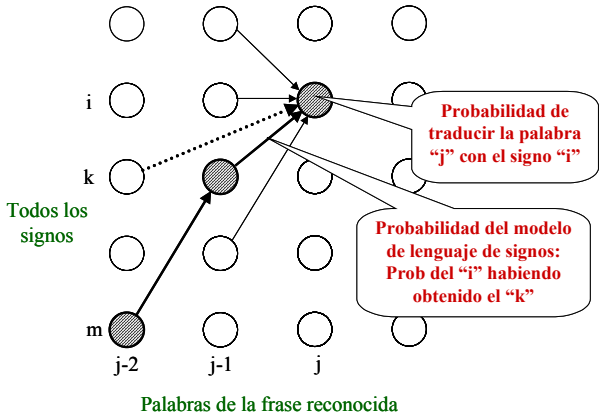

Figura 4. Ejemplo del espacio de búsqueda y la combinación de ambas probabilidades.

Ambas probabilidades se combinan en un algoritmo de programación dinámica que permite ir explorando las mejores secuencias de signos dada una frase. Para obtener la secuencia de signos final se debe realizar un backtracking sobre el espacio de búsqueda seleccionando la secuencia de signos óptima (Figura 4).

En el caso de la probabilidad de la secuencia de signos se ha utilizado un modelo de lenguaje 3-gram (la probabilidad de un signo depende de los dos anteriores) con backoff, entrenado con las secuencias de signos resultado de la traducción de las frases de entrenamiento.

La probabilidad de traducir unas palabras por un signo se ha obtenido mediante el alineamiento entre palabras y signos de las frases de entrenamiento. La probabilidad de traducción de una palabra por un signo se obtiene computando el número de veces que una palabra concreta queda alineada con un signo concreto.

Este proceso es iterativo de forma que las probabilidades estimadas en una iteración se utilizan para el alineamiento de la siguiente. Tras cada alineamiento, las probabilidades se vuelven a estimar. En la primera iteración se utiliza una tabla de traducción directa entre palabras y signos cuya traducción es unívoca. Esta primera tabla permite hacer un primer alineamiento y comenzar el proceso iterativo de estimación de probabilidades.

7. Resultados

Para evaluar el funcionamiento de este primer prototipo, se ha calculado los porcentajes de signos correctos, signos insertados (en relación con la referencia de traducción), signos borrados (según la referencia), y signos sustituidos. Para obtener estos porcentajes, la secuencia de signos resultado de la traducción se compara con la referencia (o referencias si hay varias alternativas de traducción) mediante un algoritmo de programación dinámica (o distancia de Levenshtein) considerando costes

iguales para todos los tipos de errores. Con estos porcentajes es posible calcular el porcentaje de signos con error (PSE) de forma similar a como se calcula la tasa de error en un sistema de reconocimiento de habla. En esta evaluación también se ha calculado la medida BLEU (BiLingual Evaluation Understudy). Esta medida es menos estricta que la PSE y está siendo muy utilizada en traducción automática [12]. Los resultados finales obtenidos se presentan en la Tabla 1.

Tabla 1. Resultados finales obtenidos con el sistema basados en reglas y el sistema basado en una traducción estadística

MODELO	TIPO	BLEU	PSE (%)	INS (%)	BOR (%)	SUB (%)
REGLAS	TEXTO	0.79	16.8	4.2	10.2	2.4
	VOZ	**0.64**	**25.2**	**5.9**	**16.8**	**2.5**
ESTADÍSTICA	TEXTO	0.85	15.4	6.0	7.2	2.2
	VOZ	**0.70**	**24.0**	**10.0**	**9.5**	**4.5**

Como se puede observar, el módulo de traducción basado en reglas presenta un alto porcentaje de borrados en comparación con el resto de errores. Este hecho es debido a la estrategia de las reglas: el reconocedor de voz comete errores que hace que varios patrones de palabras no aparezcan (para activar algunas de las reglas definidas) y dichas reglas no generan correctamente los signos resultado de dicha traducción.

Por otro lado la traducción estadística produce un mayor número de inserciones y sustituciones generando una gran cantidad de signos erróneos en la frase en LSE resultado de la traducción. Estos signos pueden modificar sensiblemente el significado de la frase final. En determinadas situaciones es mejor omitir dicha información que generar signos erróneos. La información omitida puede ser preguntada otra vez pero signos erróneos puede dar lugar a una confusión. Por esta razón la traducción basada en reglas puede ser preferida en determinadas aplicaciones.

8. Representación de los signos

Para la representación de los signos se presentan dos avatares. Uno de ellos, es 'VGuido', el avatar desarrollado en el marco el proyecto europeo eSIGN. En el trabajo presentado en este artículo se ha desarrollado otro avatar, centrándonos en dotar de especial detalle las partes más importantes a la hora de representar los signos: los brazos y la cabeza. Para ello se ha realizado un sistema basado en jerarquías de huesos, implementado sobre OpenGL, resultando un prototipo denominado 'Sara'. Se comentará el caso de Sara en detalle. El sistema de huesos para los brazos consta de dos esqueletos (con 17 huesos cada uno), uno para cada brazo. La cabeza es un objeto en formato .3DS (3D StudioMax), que se importa como objeto aparte. A continuación se explican en detalle cada una de estas partes del avatar (Figura 5).

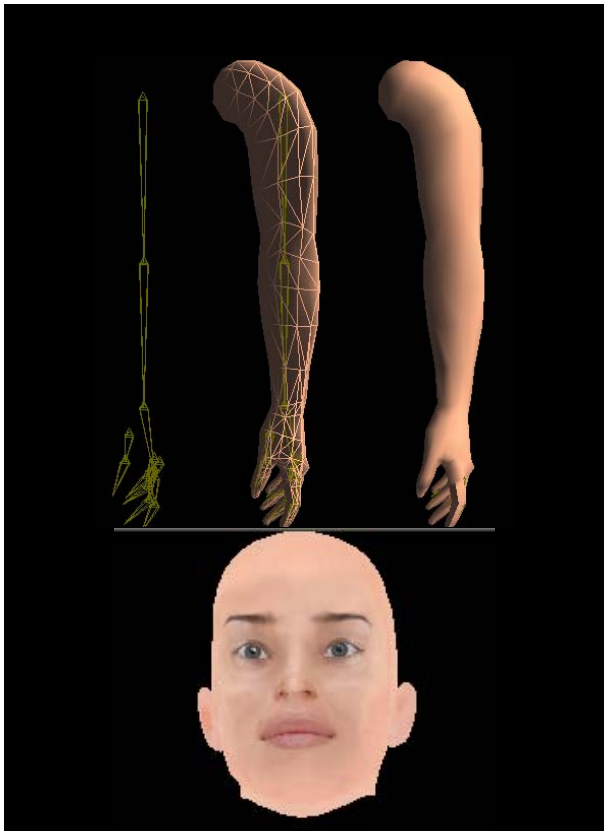

Figura 5. Sistema de representación de animaciones: sistema jerárquico de huesos, con textura asociada, y cabeza del avatar.

8.1. El sistema de brazos.

El sistema de brazos consta de dos esqueletos, uno para cada brazo. Cada esqueleto (cada brazo) tendrá un hueso padre, u origen, del cual dependerán los demás. Asimismo, cada hueso tendrá un padre, que será el hueso inmediatamente superior, y uno o varios hijos. De esta forma se asegura un movimiento realista y proporcionado. Los huesos están cubiertos con una malla, a modo de piel. Dicha malla se asocia al sistema mediante un cálculo ponderado de pesos. Cada punto de la malla estará asociado a un hueso concreto (si está claro cuál es el hueso más cercano). Es decir, se tienen dos tipos de casos: puntos de la piel que deben moverse con un hueso concreto y puntos de la piel que, por estar en zonas críticas, deben moverse según el movimiento de dos huesos (zonas de unión entre huesos). Una vez definidos los esqueletos, se describe el sistema de movimiento.

Cada hueso está definido por las traslaciones tridimensionales (x,y,z) de su punto de origen (el punto de unión con el hueso padre), y las rotaciones tridimensionales. A la hora de representar un movimiento, se ha comprobado que este queda

perfectamente definido por las rotaciones parciales de cada uno de los huesos únicamente. Cada vez que se mueve un hueso, se actualizan sus rotaciones, haciendo por extensión que se actualicen las rotaciones de los huesos hijos. Al final se adapta la malla (piel) a la nueva situación de los huesos.

8.2. La cabeza.

Es un objeto importado en el sistema en formato *.3DS. No dispone de tanta adaptabilidad como los brazos, solamente es capaz de realizar movimientos, sin expresiones. No obstante, es suficiente para lo que se quiere implementar, ya que su principal función es acompañar o servir de referente a los movimientos de los brazos.

8.3. Funcionamiento del sistema.

En el procedimiento para diseñar una animación se realizan las siguientes acciones, por este orden: creación de posiciones (parciales y/o totales), creación de animaciones.

8.3.1. Creación de posiciones.

La herramienta desarrollada ofrece la posibilidad de crear y guardar posiciones, para cargarlas posteriormente, y 'resetear' dichas posiciones. Se puede trabajar con posiciones del esqueleto completo, o posiciones únicamente de las manos. En ambos casos, el procedimiento es el mismo: se guardan en un archivo los datos referentes a las rotaciones de los huesos implicados (todo el esqueleto o solamente los huesos de las manos). De esta forma se pueden guardar posiciones concretas, posiciones comunes en la representación de la lengua de signos, para posteriormente combinarlas entre sí, formando la secuencia de signos deseada.

8.3.2. Creación de animaciones.

Para la creación de animaciones, se define un array de posiciones. Las animaciones se crearán como secuencias de posiciones. Para ello, se definen tres tipos de posiciones: 0, 1 y 2. Las posiciones por defecto serán de tipo 0, que significa que no han sido especificadas. Las posiciones tipo 1 son las que se establecen específicamente con una posición concreta. Las posiciones inicial y final serán siempre de tipo 1 por defecto. La creación de las posiciones intermedias se basará en la interpolación lineal, entre las posiciones definidas de tipo 1. Las posiciones obtenidas de esta interpolación se denominarán de tipo 2. La animación consistirá en la representación sucesiva de cada posición en pantalla. La creación de animaciones resulta sencilla habiendo dotado al sistema de utilidades como: guardar, cargar y 'resetear' posiciones, creación y borrado de interpolaciones (incluso para animaciones ya guardadas), y previsualización de resultados. Las animaciones simples pueden tener un máximo de 120 posiciones. La longitud de una animación se define según se va creando, a medida que se guardan las diferentes posiciones. La duración de un signo puede modificarse si al realizarlo queda o bien rápido o bien lento en exceso. Se dispone de dos controles: uno para duplicar el número de fotogramas y otro para reducirla a la mitad. Estas operaciones pueden realizarse tantas veces como sea preciso, siempre teniendo en cuenta que el límite superior de una animación son 120 posiciones, y que no existe límite inferior salvo el que el ojo humano pueda concluir según la calidad de la animación obtenida.

Mediante la generación de animaciones, se realiza la creación de la base de datos que permitirá representar los signos provenientes del módulo de traducción descrito anteriormente. Asimismo, se presenta la posibilidad de crear animaciones nuevas, de forma que el sistema funcione de manera independiente, al servicio de un logopeda que necesite ampliar la biblioteca de signos animados.

Para la presentación en pantalla, el sistema dispone de un generador de videos, de forma que al guardar una animación (ya sea simple o compleja) se crea automáticamente un video que puede ser utilizado como representación en cualquier sistema, dotando así a la aplicación de mayor adaptabilidad. Se puede elegir la calidad de video sin más que seleccionar el códec deseado en la lista disponible.

Para la representación de grandes secuencias de signos, se ha definido lo que se llama: animación compuesta. Una animación compuesta se formará concatenando animaciones con un número de posiciones variable (con un máximo de 120). Estas secuencias se unirán mediante pequeños fragmentos de animación, de 10-20 posiciones tipo 0, de forma que cuando se tenga definida la cadena de animaciones se pueda interpolar la secuencia, creando movimientos de unión entre los distintos signos a representar.

Para el tratamiento de las animaciones compuestas se dispone de varias herramientas, mediante las cuales se puede modificar y ajustar la secuencia, para lograr el efecto deseado. En principio el número de posiciones a insertar entre las animaciones simples que se van a concatenar, sería fijo, pudiéndose borrar y añadir posiciones en cualquier parte de la secuencia, sin más que indicar la posición en la cual se pretenden insertar las posiciones, o las posiciones entre las cuales se desea borrar.

El sistema ofrece la posibilidad de presentar en pantalla el estado de la animación simple última con la que se ha estado trabajando o la animación compleja que se está creando. De esta forma es sencillo cargar o editar pequeños fragmentos antes de añadirlos a la secuencia final. Se dispone de dos controles mediante los cuales es posible este cambio de contexto visual. Para todas estas funciones, se ofrece una interfaz gráfica completa, con todas las

opciones clasificadas de forma sencilla e intuitiva para que el usuario lo utilice fácilmente (Figura 6).

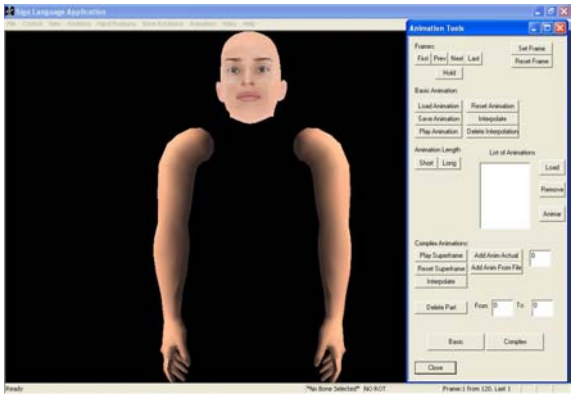

Figura 6. Interfaz gráfica de la aplicación para la representación de signos mediante el avatar 'Sara'.

Figura 7. Agente animado VGuido

El segundo avatar considerado es VGuido. VGuido es un agente animado en 3D que se ha incorporado en el sistema como un control ActiveX. Este agente animado ha sido desarrollado en el proyecto eSIGN (http://www.sign-lang.uni-hamburg.de/eSIGN/) y permite representar la secuencia de gestos concatenando diferentes animaciones asociadas a cada uno de los gestos (Figura 7).

Cada una de las animaciones asociadas a los signos a representar se definen utilizando un script en lenguaje SiGML (Signing Gesture Markup Language). SiGML es una aplicación del lenguaje XML. Mediante este lenguaje se definen aspectos relacionados con la posición de las manos, la cabeza, la velocidad de signado, el tamaño del gesto, etc. La descripción de estos aspectos está basada en la notación HamNoSys [21]. Se dispone de un entorno visual que permite generar los ficheros de tipo SiGML según el signo que se desea.

9. Conclusiones

Este artículo ha presentado los primeros experimentos de un prototipo de traducción de voz a Lengua de Signos Española (LSE) en un dominio restringido. Este dominio consiste en las frases pronunciadas por un funcionario cuando atiende a las personas que desean obtener o renovar el DNI o el pasaporte. El sistema de traducción implementado está formado por 3 bloques: un reconocedor de voz para decodificar la voz en palabras. Posteriormente un módulo de traducción de lenguaje natural genera la secuencia de signos que correspondería con la salida del reconocedor. Finalmente, un módulo con un agente animado en 3D anima la secuencia de signos resultado de la traducción.

En estos experimentos se han evaluado dos alternativas de traducción. En el caso de la traducción basada en reglas se ha obtenido un 25.2% PSE (Porcentaje de Signos Erróneos) y un BLEU (BiLingual Evaluation Understudy) igual a 0.64. La segunda alternativa ha sido una traducción estadística que combina dos probabilidades: traducción y secuencia de signos. En este caso se ha obtenido un 24.0% PSE y un BLEU de 0.70.

Como se puede observar, el módulo de traducción basado en reglas presenta un alto porcentaje de borrados en comparación con el resto de errores. Por otro lado la traducción estadística produce un mayor número de inserciones y sustituciones generando una gran cantidad de signos erróneos en la frase en LSE resultado de la traducción. En determinadas situaciones es mejor omitir dicha información que generar signos erróneos. Por esta razón la traducción basada en reglas puede ser preferida en determinadas aplicaciones aunque obtenga unos resultados ligeramente peores.

Agradecimientos

Los autores quieren agradecer al consorcio eSIGN (Essential Sign Language Information on Government Networks) el permiso para utilizar el eSIGN Editor y el avatar en 3D desarrollado por ellos (VGUIDO). Este trabajo ha sido financiado por los siguientes proyectos de investigación TINA (UPM-CAM. REF: R05/10922), ROBINT (DPI2004-07908-C02) y EDECAN (TIN2005-08660-C04).

Referencias

[1] Stokoe, W., (1960). "Sign Language structure: an outline of the visual communication systems of the American deaf". Studies in Linguistics. Buffalo.

[2] Christopoulos, C. Bonvillian, J., (1985). "Sign Language". J. of Communication Disorders, 18 1-20.

[3] Pyers J.E., (2006) "Indicating the body: Expression of body part terminology in American Sign Language". Language Sciences. Available online 4 January 2006.

[4] Hansen, B., (1975). "Varieties in Danish Sign Language". Sign Language Studies, 8: 249-256.

[5] Kyle, J., (1981). "British Sign Language". Special Education, 8: 19-23. 1981.

[6] Penn, C., Lewis, R., and Greenstein, A., (1984). “Sign Language in South Africa”. South African Disorder of Communication, 31: 6-11. 1984.
[7] Notoya, M., Suzuki, S., Furukawa, M., and Umeda, R., (1986). “Method and Acquisition of Sign Language in Profoundly Deaf Infants”. Japan Journal of Logopedics and Phoniatrics, 27: 235-243. 1986.
[8] Rodríguez, M.A. (1991). “Lenguaje de signos” Phd Dissertation. CNSE and ONCE. Madrid. Spain.
[9] Och J., H. Ney. (2002). “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation”. Annual Meeting of the Ass. ACL, Philadelphia, PA, pp. 295-302. 2002.
[10] Sumita E., Y. Akiba, T. Doi et al. (2003). “A Corpus-Centered Approach to Spoken Language Translation”. Conf. Of Ass. for Computational Linguistics (ACL) Hungary. pp171-174.
[11] Casacuberta F., E. Vidal. (2004). “Machine Translation with Inferred Stochastic Finite-State Transducers”. Comp. Linguistics, V30, n2, 205-225.
[12] Papineni K., S. Roukos, T. Ward, W.J. Zhu. (2002) “BLEU: a method for automatic evaluation of machine translation”. 40th Annual Meeting of the ACL, Philadelphia, PA, pp. 311-318. 2002.
[13] Och J., H. Ney. (2003). “A systematic comparison of various alignment models”. Computational Linguistics, Vol. 29, No. 1 pp. 19-51, 2003.
[14] Koehn P., F.J. Och D. Marcu. (2003) “Statistical Phrase-based translation”. Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, pp. 127-133, May 2003.
[15] http://www.sign-lang.uni-hamburg.de/eSIGN/
[16] S. Morrissey and A. Way. 2005. An example-based approach to translating sign language. In Workshop Example-Based Machine Translation (MT X–05), pages109–116, Phuket, Thailand, September.
[17] M. Huenerfauth. 2004. A multi-path architecture for machine translation of English text into American Sign language animation. HLT-NAACL, Boston, MA, USA.
[18] S.J. Cox, M. Lincoln, J Tryggvason, M Nakisa, M. Wells, Mand Tutt, and S Abbott. TESSA, a system to aid communication with deaf people. In ASSETS 2002, pages 205-212, Edinburgh, Scotland, 2002.
[19] J. Bungeroth and H. Ney: Statistical Sign Language Translation. In Workshop on Representation and Processing of Sign Languages, LREC 2004, 105-108.
[20] J.M. Montero: Manual del SERVIVOX. GTH-DIE. UPM, 1998.
[21] Prillwitz, S., R. Leven, H. Zienert, T. Hanke, J. Henning, et-al. (1989). Hamburg Notation System for Sign Languages – An introductory Guide. International Studies on Sign Language and the Communication of the Deaf, Volume 5. Institute of German Sign Language and Communication of the Deaf, University of Hamburg, 1989.