AVIVAVOZ: TECNOLOGÍAS PARA LA TRADUCCIÓN DE VOZ

José B. Mariño Acebal

Centro de Investigación TALP
Universidad Politécnica de Cataluña
canton@gps.tsc.upc.edu

RESUMEN

AVIVAVOZ es un proyecto de tres años dirigido a la investigación avanzada en todas las tecnologías clave que intervienen en un sistema de traducción de voz (reconocimiento, traducción y síntesis de voz).
El objetivo del proyecto es lograr avances reales en todos los componentes de un sistema de traducción de voz para alcanzar sistemas de intermediación oral entre personas en las lenguas oficiales del estado español (castellano, catalán, euskera y gallego) entre sí y entre el castellano y el inglés.
El proyecto aborda el avance y la integración de las tres tecnologías implicadas. En reconocimiento de voz se desarrollará un sistema robusto, en un dominio amplio de aplicación (noticias de radiotelevisión y sesiones parlamentarias) y gran vocabulario. En traducción se avanzará en el desarrollo de técnicas estadísticas de traducción incluyendo la incorporación de distintas fuentes de conocimiento lingüístico (detección de eventos, análisis sintáctico y semántico). En síntesis de voz se generarán nuevos modelos acústicos y prosódicos para generar voz expresiva. La interacción e integración entre las tres tecnologías constituye el cuarto de los problemas abordados en este proyecto.
Las tecnologías desarrolladas en el proyecto participarán en campañas de evaluación competitivas internacionales y los avances científicos y tecnológicos se mostrarán mediante un demostrador de doblaje automático de informativos y de discursos parlamentarios.

1. INTRODUCCIÓN[1]

Puede decirse que el interés por la aplicación del computador a la traducción automática surge inmediatamente con el desarrollo de los primeros computadores tras la segunda guerra mundial. Sin embargo, tras muchos esfuerzos de investigación en EEUU la escasa capacidad de los computadores de la época no permite alcanzar resultados de interés práctico y la actividad se desvanece.

Durante los años 70 y 80 el interés en la traducción automática se desplaza a las sociedades multiculturales, como Canadá y Europa. Resultado del esfuerzo realizado son los sistemas Météo y Systran.

A partir de los años centrales de la década de los noventa, la confluencia de la rápida evolución de las prestaciones de los computadores, el desarrollo de Internet, la creciente globalización mundial con el correspondiente incremento de los intercambios interculturales, incentivó de nuevo el interés en la traducción automática. Adicionalmente, las tecnologías del reconocimiento y síntesis de voz habían alcanzado el suficiente grado de madurez para sugerir la viabilidad de la traducción de voz (speech-to-speech translation) mediante un sistema resultante de la concatenación de un reconocedor de voz (convertidor de voz a texto en la lengua origen), un traductor de texto (convertidor de texto entre las lenguas origen y destino) y sintetizador de voz (convertidor de texto a voz en la lengua destino).

El interés social y económico de las potenciales aplicaciones de la nueva tecnología constituyeron un fuerte acicate para atraer la atención de la comunidad investigadora. Entre dichas aplicaciones podemos citar:

- La interpretación simultánea (en congresos, conferencias, etc.)
- La intermediación entre personas (conferencias telefónicas, interacciones personales, etc.)
- Subtitulado de video (y TV).
- Doblaje en diferentes lenguas de audio y video (y TV).
- Interpretación en tele/video conferencias multilingües.
- Indexado automático multilingüe de material multimedia.
- Traducción de mensajes de voz.

Los proyectos C-STAR [1], Verbmobil [2] y EUTRANS [2] supusieron el punto seminal de la tecnología dedicada a la traducción de voz. En particular, el proyecto Verbmobil, financiado íntegramente por el gobierno alemán, tuvo una especial significación al permitir comparar el grado éxito en la tarea de las diversas opciones aplicadas en la traducción. Así, se atrajo la atención sobre la traducción automática estadística (TAE), que por aquel entonces comenzaba su andadura a partir del trabajo de P. Brown [4] y sus colegas en el Watson Research Center de IBM.

[1] Este trabajo ha sido subvencionado por el Gobierno Español mediante el proyecto coordinado TEC2006-13694-C03.

Los proyectos mencionados estaban dedicados a la traducción de elocuciones que se producen en tareas de contenido semántico reducido. Por ejemplo, C-STAR se dedicaba a las necesidades de un turista, Verbmobil se centraba en la concertación de una cita y EUTRANS abordaba las necesidades del huésped de un hotel. El nuevo siglo, con el viento a favor de los avances conseguidos, nos trae proyectos más ambiciosos como TC-STAR [5] y GALE [6], por ejemplo, en los que la tarea a abordar implica vocabularios no limitados y contenidos semánticos no restringidos. En particular, TC-STAR está dedicado a la traducción de las intervenciones en el parlamento europeo entre el inglés y el castellano (y viceversa), y de noticias de radiodifusión de chino a inglés.

En los últimos años, la comunidad dedicada al desarrollo de la tecnología de traducción de voz ha crecido rápidamente, lo que ha motivado el interés de la realización de campañas anuales de evaluación que permitan comparar en un marco común los diversos planteamientos seguidos. A este respecto, cabe destacar las evaluaciones organizadas por el consorcio C-STAR y cuyos resultados son motivo de estudio en el International Workshop on Spoken Language Translation [7], y las campañas realizadas por el National Institute of Standards and Technology (NIST) norteamericano [8].

El proyecto AVIVAVOZ [9] se plantea situar las lenguas oficiales del estado español al nivel tecnológico del estado del arte. En particular, toma como referencia los avances alcanzados en TC-STAR.

2. LAS TECNOLIGÍAS IMPLICADAS

AVIVAVOZ es un proyecto dedicado a la investigación avanzada de las tecnologías centrales para la traducción automática del habla (TAH): reconocimiento de voz, traducción estadística y síntesis de voz.

2.1. Reconocimiento de voz

Es indudable que disponer de un sistema de reconocimiento de habla (RAH) preciso y competente es uno de los puntos clave para abordar un sistema de traducción automática voz a voz. Los sistemas de reconocimiento actuales son capaces de lograr tasas de reconocimiento extremadamente altas en aplicaciones complejas (habla continua y grandes vocabularios), pero sólo con condiciones controladas, es decir, cuando la grabación es limpia y contiene voz previamente segmentada, y además tanto el locutor, como el vocabulario y el tipo de lenguaje son conocidos y están bien modelados. Cuando una o varias de estas condiciones no se cumplen, los reconocedores de voz reducen sus prestaciones de forma importante. Este será, en general, el caso de los sistemas de traducción automática. Por un lado, y dependiendo de la aplicación, cabe esperar unas condiciones de grabación distantes de las ideales, pero además el tipo de habla, aunque no totalmente espontánea, estará bastante alejada del habla planeada o del dictado, tanto en cuanto al cuidado en la pronunciación como en cuanto al vocabulario y al lenguaje.

La aplicación que se plantea es, por tanto, de una dificultad similar a las tareas más complejas que se abordan actualmente en reconocimiento de habla: transcripción automática de programas de noticias, transcripción de conversaciones telefónicas y transcripción de reuniones de trabajo. Aunque los avances en los últimos años han sido importantes, los sistemas de reconocimiento más avanzados como el Byblos BBN en EEUU, el sistema del laboratorio LIMSI-CNRS en Francia, o el CU-HTK de la universidad de Cambridge en Reino Unido, aún no consiguen prestaciones totalmente satisfactorias en estas tareas. En transcripción de programas de noticias, el mejor resultado obtenido en la actualidad está en torno al 10% de WER (“word error rate”), mientras que para transcripción de conversaciones telefónicas la tasa de error se ve incrementada hasta el 15-20%. Estos resultados se logran con cantidades ingentes de material de entrenamiento y en tiempos de ejecución varias veces por encima de tiempo real, lo que dificulta su aplicación en problemas reales y en idiomas minoritarios.

Para lograr estas prestaciones, los problemas que tienen que resolverse (aparte de la recogida de datos) tienen que ver con el tratamiento adecuado de la alta variabilidad acústica, de lenguaje y vocabulario presentes en las interacciones entre humanos. En cuanto al modelado acústico, los problemas aparecen cuando existe desajuste entre el material de reconocimiento y el de entrenamiento. Este desajuste puede estar producido por variación del ruido ambiental o por la variabilidad interlocutor. Respecto al ruido ambiental, los reconocedores incluyen parametrizaciones que incorporan parte de las características del oído humano en cuanto a robustez frente al ruido (parametrización MFCC o PLP). Otros autores proponen la utilización de métodos de combinación de parámetros aplicando algoritmos de fusión de datos. Respecto a la variabilidad interlocutor, la normalización de la longitud del tracto vocal (VTLN) y el entrenamiento adaptativo de locutor son los sistemas más utilizados para evitar su influencia.

En cuanto al modelado lingüístico, las técnicas más empleadas para RAH de grandes vocabularios son los modelos estadísticos tipo N-grama. Existen herramientas de dominio público que permiten entrenar dichos modelos; una de las más conocidas es “SRILM - The SRI Language Modeling Toolkit” de la firma SRI. A partir de esta herramienta se pueden explorar técnicas de selección de material textual apropiado a la tarea a reconocer. Estamos hablando entonces de reducir el desajuste entre entrenamiento y reconocimiento mediante la combinación de múltiples modelos de lenguaje cada uno de ellos especializado en un “tema” diferente. Esta vía de investigación es muy interesante pues va en la línea general del proyecto de

incorporación temprana de las fuentes de conocimiento en contraposición del uso ingente de datos (voz y texto). Redundando en lo anterior, una de las estrategias más frecuentemente adoptadas para hacer viable la etapa de decodificación es el empleo de un esquema multipase. En el primer pase de reconocimiento se emplean modelos acústico y de lenguaje muy generales. La salida de este primer pase se emplea para seleccionar y refinar los nuevos modelos a usar en el segundo pase. Este procedimiento se puede realizar varias veces de forma sucesiva, utilizando en cada pase nuevas o adaptadas fuentes de conocimiento.

Debido a la naturaleza de la aplicación que nos ocupa, es de esperar que el tipo de habla al que deba enfrentarse el reconocedor sea en gran parte de tipo espontáneo. El reconocimiento de habla espontánea presenta una problemática específica y compleja, que conlleva unas prestaciones de los reconocedores mucho más reducidas que en aplicaciones de dictado. La razón principal es que el habla espontánea incluye multitud de "palabras relleno", vacilaciones, repeticiones, palabras cortadas, frases sin terminar y otro tipo de disfluencias. Estos fenómenos son muy difíciles de modelar tanto desde el punto de vista léxico, como lingüístico. Los modelos de lenguaje que se entrenan a partir de textos escritos fallan de forma evidente al ser aplicados a este tipo de tareas, mientras que la construcción de modelos de lenguaje específicos resulta difícil debido a la escasez de material de la que se dispone. La incorporación de información lingüística de mayor alcance que la que actualmente se hace servir es una opción a considerar. Esto aplica tanto al modelado de lenguaje (donde sólo se está usando n-gramas) como al modelado prosódico (donde características a largo plazo deberían ser más beneficiosas que las actuales a corto plazo, aunque ello implique un aumento en la complejidad del algoritmo de búsqueda).

La traducción debe enfrentarse con los errores de reconocimiento. Medidas de confianza adjuntas a las palabras en una lista con las N-mejores hipótesis (o, alternativamente, un grafo) pueden ser una aportación para paliar los efectos de un reconocimiento defectuoso. Por ello, en este proyecto se continuarán los trabajos para el desarrollo de nuevas medidas de confianza basadas en test de hipótesis Bayesiano que se incluirán en las nuevas versiones del motor de reconocimiento. Además, la traducción requiere información adicional a la mera transcripción ortográfica del mensaje oral. Por ejemplo, se necesita información sobre puntuación que puede ser capturada mediante modelos de lenguaje adecuados y/o algoritmos mejorados de segmentación acústica. Ambos problemas serán considerados en este proyecto.

En la literatura dedicada al reconocimiento de voz, la detección de "metadatos" está adquiriendo relevancia. Hasta ahora, los sistemas de reconocimiento proporcionaban solamente transcripción de la voz. Cuando el mensaje debe ser reproducido en otra lengua con una voz similar a la original, se hace necesaria información adicional, como la identidad del hablante o la indicación de cambio de hablante. Esta información debe ser proporcionada al sintetizador de voz. Entre otros parámetros de interés, podemos mencionar: velocidad de articulación, pausas y otras características prosódicas (disfluencias, "palabras relleno", etc.). En este proyecto, se incluyen tareas dedicadas a la extracción de estos "metadatos" de la señal de voz y su utilización en los módulos de traducción y síntesis.

A los problemas anteriores, se añade el interés en el desarrollo de sistemas de reconocimiento multilingües sobre los que nos centraremos en dos aspectos: portabilidad de tecnología entre idiomas y funcionamiento en condiciones de bilingüismo.

2.2. Traducción estadística

En tan solo diez años, los métodos de traducción automática estadística (TAE) han alcanzado prestaciones similares a las de los métodos basados en conocimiento, los cuales han estado evolucionando por más de medio siglo. Sin embargo, a pesar de los progresos actuales, la tecnología de la TA está todavía muy lejos de alcanzar niveles de calidad y prestaciones satisfactorios.

Entre las principales limitaciones que exhiben los sistemas de TA actuales, las más importantes son su limitación a trabajar en dominios restringidos, su incapacidad para manejar vocabularios muy grandes, y sus aún relativamente altas tasas de error. Aunque los expertos opinan que un verdadero avance en la tecnología de la TA sólo será posible mediante la combinación de la fuerza bruta (métodos estadísticos) con fuentes de información lingüística (métodos basados en conocimiento), muchos de los esfuerzos recientes en esta dirección han fracasado en proporcionar mejoras significativas respecto a los sistemas de TA existentes. Por esta razón, hoy por hoy, la TA continúa siendo uno de los dolores de cabeza más interesantes de la Inteligencia Artificial (IA), y un gran reto para académicos e investigadores.

Desde el punto de vista de su aplicación práctica, la TA se puede dividir en dos problemas específicos: la traducción automática del lenguaje escrito (TALE) y la traducción automática del lenguaje hablado (TAH). El problema de la traducción del habla (TAH) presenta, en adición a las complejidades específicas de la TA, dos tipos de problemas adicionales: en primer lugar, aquellos relacionados con la naturaleza del lenguaje oral, como por ejemplo la espontaneidad y la mayor libertad de las estructuras sintácticas; y en segundo lugar, aquellos relacionados con el estado actual del arte en el reconocimiento automático del habla (RAH), como por ejemplo los errores de reconocimiento. Por estas razones, se ha prestado especial atención al problema de la integración de las tecnologías de RAH y TA en los últimos años.

En el caso de la TALE, a diferencia de la TAH, el lenguaje escrito es mucho más controlado en lo relativo

a contenido gramatical de las oraciones. Esto permite explotar una mayor cantidad de información lingüística partiendo desde el nivel léxico hasta los niveles sintáctico y semántico. Recientemente, algunos autores han presentado propuestas para usar información sintáctica para construir modelos de traducción, reportando resultados prometedores en la mejora de la calidad lingüística de las traducciones.

Este proyecto considera el problema de TA en su variante TAH y se concentra en tareas de traducción entre las cuatro lenguas del estado español, así como entre el castellano y el inglés; en función de lo que permitan los recursos bilingües disponibles. Aunque el problema de TA será principalmente abordado desde el paradigma estadístico, también se dedicará un esfuerzo importante a la incorporación, uso y adaptación de las herramientas y recursos lingüísticos disponibles. Así mismo, se prestará especial atención a la integración de los sistemas de RAH y TA mediante el uso de modelos de lenguaje y traducción basados en n-gramas. Adicionalmente, se trabajará en el desarrollo de algoritmos novedosos y métodos para la inclusión de información lingüística con el objeto de mejorar el estado actual del arte en TAE tanto desde el punto de vista de la eficiencia computacional como de las prestaciones en términos de calidad de traducción, tamaño del vocabulario y restricciones de dominio.

2.3. Síntesis de voz

En la última década se ha producido una significativa mejora en la calidad de la voz sintética. Sin embargo, esta tecnología está aún restringida a aplicaciones donde es aceptable una limitada calidad de la voz, o donde el dominio de aplicación es muy específico. La síntesis de voz se utiliza actualmente cuando las comunicaciones entre persona y máquina serían muy difíciles o incluso imposibles sin la ayuda de estos sistemas. El objetivo final de estos sistemas es obtener voces naturales capaces de expresar cualquier estilo, humor, acento u otra característica de los hablantes humanos. Un escenario en el que se utilizan estas aplicaciones es la traducción voz-voz, donde la voz sintética traducida podría reproducir las características del hablante original. Otro aspecto importante a considerar es la lengua utilizada en el proceso de comunicación. Cada vez es más necesario el acceso multilingüe a distintos dispositivos y sistemas. Esto es especialmente relevante en áreas donde hay más de una lengua oficial. Para conseguir estos objetivos, en este proyecto se desarrollarán nuevas técnicas para la generación de prosodia y la producción de voz, en un entorno multilingüe.

Las técnicas de síntesis por concatenación de unidades han permitido la mejora de la calidad acústica de la señal sintética. En estas técnicas, segmentos de señal extraídos de grabaciones previas realizadas con un locutor, se concatenan para producir la señal de salida deseada. La prosodia de la señal de salida debe ser calculada de acuerdo a los modelos prosódicos adecuados. Para generar una transición natural entre los segmentos de voz y conseguir la prosodia deseada, es necesario manipularlos adecuadamente. Para manipular la señal de voz con alta calidad, se han utilizado distintos algoritmos como PSOLA, MBROLA, HNM y modelos sinusoidales. Con estas técnicas, la calidad de la señal sintética de salida disminuye drásticamente si los segmentos manipulados tienen unas características muy alejadas de las de los originales. Sin embargo, son muy apropiadas en sistemas con poca memoria y baja capacidad de cálculo, como los dispositivos portátiles. Obtener mejores modelos de la producción de la voz mejoraría la calidad final de todos estos sistemas.

La solución actual al problema de la distorsión debida a la manipulación de las unidades de voz es grabar tantos segmentos de voz como sea posible, con la esperanza de que a la hora de la síntesis la manipulación se reduzca al mínimo. Para ello se diseña un corpus que contenga todos los sonidos en todos los contextos fonéticos y con todos los contornos prosódicos que puedan ser necesarios en el momento de la síntesis. En muchos casos, no es necesario manipular significativamente la voz original ya que los segmentos seleccionados tienen unas características prosódicas muy parecidas a las buscadas para la señal sintética, obteniéndose una calidad de la voz sintética excelente. Por otro lado, si es necesaria una manipulación significativa, la voz sintética resultante suena muy poco natural y puede llegar a ser incomprensible. En el diseño de un sistema de síntesis basada en corpus hay tres problemas principales. El primero es el diseño del corpus de grabación: cuanto más grande sea, mayor cobertura tendrán las unidades grabadas de modo que se reduce la probabilidad de introducir distorsión por manipulación de las unidades. Sin embargo, muchos aspectos del lenguaje están caracterizados por un gran número de fenómenos raros, lo que implica que es necesaria una base de datos enorme. Se han desarrollado técnicas para obtener bases de datos óptimas, con resultados muy prometedores, pero es necesario profundizar más. El segundo problema es la definición del coste de concatenación de las unidades. En el momento de la concatenación hay que seleccionar las mejores unidades, lo que implica definir una distancia entre las unidades seleccionadas y las unidades objetivo y una distancia entre las características de las unidades. Para cada una de las características utilizadas es necesario definir distancias perceptuales. Finalmente, encontrar la mejor combinación de unidades requiere una búsqueda por toda la base de datos, lo cual implica una gran capacidad de cálculo. Como consecuencia, es necesario utilizar algoritmos de búsqueda rápida.

De los párrafos anteriores se deduce que obtener una nueva voz con alta calidad es una tarea que tiene un coste muy elevado que requiere la selección del locutor y la grabación del corpus. La transformación de la voz del locutor mediante manipulación de sus características es una técnica muy interesante que ya está ofreciendo

resultados prometedores. En las aplicaciones de traducción voz-voz, la reproducción de las características del hablante original (calidad de la voz, estilo de habla, emoción, etc.) es un interesante objetivo.

Uno de los mayores retos que existen en conversión texto a voz es dotar a la voz sintética de la suficiente naturalidad, que se incrementaría al poder expresar emociones. Para producir habla sintética emocional es preciso contar con una buena base de datos de voz emocional que permita el estudio y la caracterización acústica de las emociones. Antes de ser utilizadas, estas bases de datos deben ser evaluadas, ya que la emoción emitida y la percibida no siempre coinciden. La síntesis del habla emocional se realiza principalmente utilizando técnicas basadas en corpus o por regla, aunque hay otras aproximaciones, como usar una representación fonológica intermedia o realizar *morphing* prosódico. Para poder considerar las diferentes emociones en el texto de entrada al sistema, hay que contemplar una interfaz entre la aplicación y el sistema de conversión de texto a voz. El lenguaje de marcado desarrolla esta interface. Actualmente se están definiendo los lenguajes de marcado para cumplir esta función, todos ellos basados en el estándar XML. Entre los lenguajes con aplicación a la síntesis de voz emocional se incluyen el lenguaje de presentación multimodal (MPML, *Multimodal Presentation Markup Language*) y la especificación del lenguaje de marcado para humanos virtuales (VHML, *Virtual Human Markup Language*). Esta última incluye un sublenguaje específico para el marcado de emociones (EML *Emotion Markup Language*).

3. EL CONSORCIO INVESTIGADOR

El proyecto AVIVAVOZ ha sido propuesto por un consorcio formado por tres equipos investigadores. Los grupos participantes son el grupo de Procesado de Voz del "Centro de Investigación de Tecnologías y Aplicaciones de la Tecnología del Lenguaje y el Habla" (TALP) de la Universidad Politécnica de Cataluña, el Grupo de Teoría de la Señal (GTS) del Departamento TSC de la Universidad de Vigo y el Grupo de Investigación Aholab (AHOLab) de la Universidad del País Vasco. Los tres grupos tienen experiencia previa en participación y colaboración conjunta en proyectos coordinados financiados por CICYT y por empresas y en la actualidad participan en una red de excelencia internacional [ECESS] de síntesis de voz.

TALP [10] aporta experiencia en los tres campos: reconocimiento de voz con sistemas competitivos en aplicaciones telefónicas y un sistema de reconocimiento de habla continua, ambos en catalán, castellano e inglés, traducción de habla en catalán, castellano e inglés y síntesis de voz en castellano y catalán. TALP coordinará las actividades de traducción e integración de tecnologías.

GTS [11] aporta experiencia en reconocimiento de habla continua en una tarea compleja como es noticias radiofónicas en gallego y castellano, que complementa perfectamente el sistema de TALP y una sólida experiencia en síntesis de voz en ambos idiomas. GTS coordinará las tareas de reconocimiento de voz.

Finalmente, AHOLab [12] es un grupo experto en síntesis de voz y en particular en prosodia. Su sistema de síntesis trabaja en euskera y castellano. También ha desarrollado un sistema de reconocimiento en euskera. AHOLab coordinará las actividades de síntesis de voz.

4. OBJETIVOS DEL PROYECTO

El objetivo del proyecto es lograr avances reales en todos los componentes de un sistema de traducción de voz para alcanzar sistemas de intermediación oral entre personas en las lenguas oficiales del estado español (castellano, catalán, euskera y gallego) entre sí y entre el castellano y el inglés. En reconocimiento de voz, se desarrollará un sistema para grandes vocabularios y dominios de aplicación amplios (noticias de radiodifusión y discursos parlamentarios). Para la traducción de habla, se construirá un sistema estadístico basado en corpus al que se incorporará conocimiento lingüístico (detección de eventos, análisis sintáctico, etc.). Para la síntesis de voz, se investigarán modelos acústicos y prosódicos nuevos para la generación de voz expresiva. Por último se abordará la integración de las tecnologías anteriores y se desarrollará un demostrador.

4.1. Reconocimiento de voz

El objetivo general es aumentar las prestaciones de los sistemas de reconocimiento automático de habla de los distintos grupos de investigación de este proyecto, y lograr su integración en un sistema de traducción de voz-a-voz. Desgranándolo según actividades:

- Desarrollo de una nueva arquitectura de reconocimiento flexible, con múltiples capas, en la que la decodificación acústico-fonémica se realice de forma independiente de la decodificación de palabras.
- Desarrollo de un módulo de segmentación de audio en trozos lingüísticamente homogéneos.
- Desarrollo de nuevas medidas de confianza basadas en la información proporcionada por la arquitectura anterior.
- Mejora del modelado acústico mediante la aplicación de nuevas técnicas de adaptación al locutor y al canal.
- Mejora del modelado acústico en entornos multilingües para reducir el conjunto total de modelos acústicos.
- Desarrollo de modelos de lenguaje mediante el empleo de algoritmos de adaptación y aprendizaje.
- Tratamiento específico del habla espontánea
- Evaluación del sistema de reconocimiento de grandes vocabularios y habla continua.

4.2. Traducción estadística

El objetivo principal es desarrollar sistemas de traducción automática que incorporen información morfo-sintáctica y semántica. Este objetivo principal se puede subdividir en los siguientes objetivos específicos:

- Preprocesado y edición final de colecciones de datos bilingües de amplio dominio y anotados con información lingüística relevante, para ser usados en el entrenamiento de los sistemas de traducción automática estadística.
- Desarrollo de un sistema base para la traducción automática del habla y su integración eficiente con el sistema de reconocimiento automático.
- Desarrollo de algoritmos y estrategias de búsqueda eficientes que permitan la implementación en tiempo real de los sistemas de traducción automática estadística.
- Desarrollo de sistemas de traducción específicos para las cuatro lenguas oficiales en España.
- Demostración de las prestaciones de los sistemas desarrollados mediante la realización de tareas específicas de traducción y la comparación con otros sistemas de traducción automática.

4.3. Síntesis de voz

El objetivo principal del proyecto en el campo de la síntesis de voz es la obtención de sistemas capaces de generar voces con un alto grado de naturalidad en un entorno multilingüe (inglés, castellano, catalán, euskera y gallego). Estos sistemas deben expresar diferentes estilos de habla, acentos y en general parámetros de calidad de la voz especificados en un texto de entrada. Para lograr este objetivo, se desarrollarán los siguientes objetivos parciales:

- Interpretación adecuada del texto de entrada. Una interpretación incorrecta causada por una entrada anómala (tal como ausencia o incorrecta situación de signos de puntuación, etc.) pueden generar una salida ininteligible. De la misma forma, la transcripción de grafema a alófono debe ser precisa si se desea obtener el mensaje acústico adecuado a la entrada. En este proyecto se investigarán algoritmos de procesado lingüístico y técnicas basadas en datos que mejoren los sistemas actuales.
- Generación de voz expresiva en diferentes estilos de habla, voces e idiomas. Para obtener voz sintética expresiva con un alto grado de calidad es necesario elaborar nuevos modelos prosódicos. En este sentido se investigarán técnicas para inferir la prosodia del texto de entrada, y transferida correctamente a las unidades de síntesis, así como su relación con los parámetros acústicos relevantes.
- Algoritmos de generación manipulación de voz. Se necesitan nuevos modelos de producción de voz que permitan realizar modificaciones de las características prosódicas y espectrales de la señal sin introducir una degradación significativa. Estos modelos pueden ser utilizados para modificar las características de la voz así como para caracterizar al locutor, como es el caso de una tarea de transformación de voz. Los mismos modelos pueden utilizarse también para comprimir la señal de voz y reducir la cantidad de memoria utilizada.
- Estándares, evaluación y sistemas de referencia. Con objeto de evaluar la validez de las nuevas técnicas desarrolladas es necesario definir sistemas de referencia. Asimismo, para permitir el intercambio de módulos y la evaluación comparativa del comportamiento individual de cada módulo, se propone definir interfaces estándares de conexión de módulos.

4.4. Integración de tecnologías

En esta actividad se integran los sistemas de reconocimiento, traducción y síntesis a fin de construir el sistema completo de traducción del habla. El logro de este objetivo general implica el cumplimiento de los sub-objetivos siguientes:

- Desarrollo de recursos bilingües para los pares Castellano – Inglés/Catalán/Euskera/Gallego, incluyendo textos paralelos y herramientas de procesado de lenguaje.
- Realización de una arquitectura web para la comunicación entre los diversos subsistemas.
- Definición y realización de interfaces eficientes entre los distintos subsistemas que permitan mejorar las prestaciones del sistema completo.
- Realización de herramientas para el doblaje automático de video como demostrador de las tecnologías desarrolladas.

5. AGRADECIMIENTOS

Este trabajo se basa en la memoria técnica del proyecto AVIVAVOZ. En consecuencia, debe reconocerse (y agradecerse) a todos los miembros del equipo investigador su contribución.

6. REFERENCIAS

[1] http://www.c-star.org/
[2] http://verbmovil.dfki.de/verbmobil/overview-us.html
[3] http://www.cordis.lu/espirit/src/30268.htm
[4] P. F. Brown et al., “The mathematics of statistical machine translation: parameter estimation”, Computational Linguistics, 19 (2), 1993.
[5] http://www.tc-star.org
[6] http://www.darpa.mil/ipto/programs/gale/index.htm
[7] http://www.slt.atr.jp/IWSLT2006/
[8] http://www.nist.gov/speech/tests/mt/
[9] http://gps-tsc.upc.es/veu/avivavoz
[10] http://gps-tsc.upc.es/veu/
[11] http://www.gts.tsc.uvigo.es/web/index.php
[12] https://bips.bi.ehu.es/aholab/