Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 169-175 recibido 29-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Sistema de Conversión Texto a Voz de Código Abierto Para Lenguas Ibéricas

Open-Source Text to Speech Synthesis System for Iberian Languages

Agustin Alonso[1], Iñaki Sainz[1], Daniel Erro[1,2], Eva Navas[1], Inma Hernaez[1]
[1]AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
[2]Basque Foundation for Science (IKERBASQUE), Bilbao, Spain
{agustin,inaki,derro,eva,inma}@aholab.ehu.es

Resumen: Este artículo presenta un conversor texto a voz basado en síntesis estadística que por primera vez permite disponer en un único sistema de las cuatro lenguas oficiales en España además del inglés. Tomando como punto de partida el sistema AhoTTS existente para el castellano y el euskera, se le han añadido funcionalidades para incluir el catalán, el gallego y el inglés utilizando módulos disponibles en código abierto. El sistema resultante, denominado AhoTTS multilingüe, ha sido liberado en código abierto y ya está siendo utilizado en aplicaciones reales.
Palabras clave: Texto a Voz, Multilingüismo, Herramienta Software, Síntesis Estadística, Código Abierto

Abstract: This paper presents a text-to-speech system based on statistical synthesis which, for the first time, allows generating speech in any of the four official languages of Spain as well as English. Using the AhoTTS system already developed for Spanish and Basque as a starting point, we have added support for Catalan, Galician and English using the code of available open-source modules. The resulting system, named multilingual AhoTTS, has also been released as open-source and it is already being used in real applications.
Keywords: Text-to-Speech, Multilingualism, Software Tool, Statistical Synthesis, Open Source

1 Introducción

Un sistema de conversión texto a voz (CTV) es un sistema que convierte una entrada de texto en una salida en forma de señal de audio cuyo contenido se corresponde con el mensaje del texto de entrada.

Los sistemas CTV actuales pueden descomponerse en dos módulos tal y como muestra la Figura 1. El primero, fuertemente dependiente del idioma, toma como entrada el texto a sintetizar y genera a su salida etiquetas que describen lingüística, fonética y prosódicamente dicha entrada. Estas etiquetas alimentan el segundo módulo, el motor de síntesis que en sí mismo es independiente del idioma. Este módulo emplea voces entrenadas a partir de bases de datos que son dependientes del idioma para sintetizar la señal de voz de salida en función de las etiquetas de entrada. Respecto a los métodos de síntesis, las tecnologías más empleadas actualmente son dos: la selección y concatenación de unidades y la síntesis estadística paramétrica. La selección de unidades (Hunt y Black, 1996) consiste en generar la señal de voz concatenando segmentos de voz real. La síntesis estadística consiste en reconstruir la señal a partir de parámetros acústicos extraídos de modelos matemáticos previamente entrenados con señales de voz real (Zen et al., 2009). También se han desarrollado sistemas híbridos que combinan ambas técnicas (Ling et al., 2007).

ISSN 1135-5948

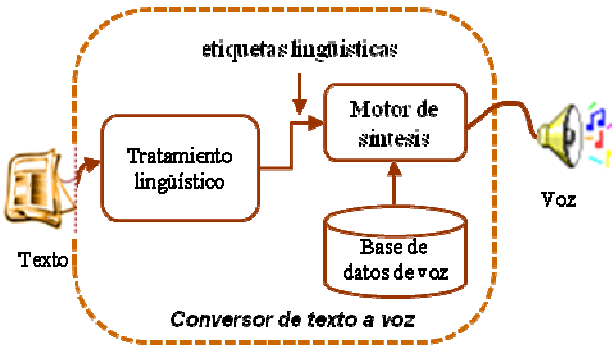

Figura 1: Estructura genérica de un sistema de conversión de texto a voz

Actualmente existen diversos conversores en código abierto desarrollados por equipos de investigación universitarios de distintas universidades para las diferentes lenguas oficiales de España que emplean múltiples métodos tanto en el procesado lingüístico como en la síntesis. Entre ellos destacan para los idiomas castellano y euskera AhoTTS[1], para el castellano y el gallego Cotovia[2], y para el catalán Festcat[3].

Aunque existen trabajos realizados para disponer en un único CTV de todas las lenguas oficiales (Rodríguez, Escalada y Torre, 1998) se trata de sistemas propietarios no disponibles libremente. La necesidad de disponer de un CTV multilingüe surge en el contexto del proyecto TVSocial[4], en el que se pretendía crear una plataforma de difusión de contenidos de televisión con bajo coste, utilizando para ello un CTV multilingüe de las características mencionadas.

En este artículo se explica el procedimiento que se ha seguido para la creación de un conversor de texto a voz de código abierto con la posibilidad de elegir entre las cuatro lenguas oficiales de España más el inglés La base de la que se ha partido es el sistema AhoTTS desarrollado por el grupo Aholab de la UPV/EHU.

Primero, en la sección 2, se describen las características básicas del sistema AhoTTS. A continuación, en la sección 3, se explican los pasos seguidos para integrar módulos de procesado lingüístico y voces nuevas para los idiomas no disponibles inicialmente. Finalmente en la sección 4 se resumen las conclusiones y se mencionan las líneas futuras de trabajo que se tiene pensado seguir.

2 *Sistema AhoTTS*

El sistema AhoTTS es la plataforma de conversión texto a voz de Aholab que lleva siendo desarrollada desde 1992. Programado en C/C++, es un sistema modular, multiplataforma y multilingüe. Inicialmente los idiomas disponibles eran el castellano y el euskera, aunque el esfuerzo de investigación se ha centrado principalmente en el segundo (Hernaez, 1995) (Navas et al., 2002a) (Navas et al., 2002b) (Navas, 2003).

AhoTTS integra la práctica totalidad de las tecnologías actuales de síntesis. Así, permite sintetizar la señal de voz empleando tanto la selección y concatenación de unidades (Sainz et al., 2008) como la síntesis estadística paramétrica (Erro et al, 2010). Para este segundo caso, Aholab ha desarrollado tecnología propia de parametrización y reconstrucción de señales (Erro et al, 2011). También se han hecho experimentos utilizando tecnologías híbridas (Sainz et al., 2011) con resultados muy satisfactorios.

De entre estas tecnologías de síntesis, la estadística presenta varias ventajas prácticas respecto al resto (Zen et al, 2009):

- Produce una voz de características más estables y con mayor inteligibilidad, especialmente cuando las bases de datos son pequeñas.
- El entrenamiento de la voz es automático, robusto y no requiere el ajuste manual de múltiples parámetros.
- El tamaño en disco de las voces generadas es menor, lo cual facilita su almacenamiento e integración en sistemas embebidos o con recursos limitados.
- Es fácil modificar las características acústicas de la voz en tiempo de síntesis y da mayor flexibilidad para generar nuevas voces mediante técnicas de adaptación, interpolación, etc.

Debido a estas ventajas se ha optado por este método para la inclusión de los nuevos idiomas en AhoTTS.

La calidad de AhoTTS viene avalada por los excelentes resultados cosechados en

[1] http://sourceforge.net/projects/ahotts/
[2] http://sourceforge.net/projects/cotovia/
[3] http://festcat.talp.cat/download.php
[4] TVSocial ETORGAI Televisión Social - Low Cost Telebista (ER-2010/00003)

evaluaciones competitivas tanto a nivel nacional como internacional. En la campaña Albayzin 2010 obtuvo el primer puesto (Sainz et al., 2010) y en la edición de 2012 obtuvo la mejor valoración para síntesis de voz neutra (Sainz et al., 2012a). En la campaña Blizzard Challenge 2011 obtuvo el 5º puesto a nivel mundial. (Sainz et al, 2011). Recientemente obtuvo los mejores resultados en varias categorías del Hurricane Challenge (Erro et al., 2013), concretamente en aquellas en las que se evaluaba la inteligibilidad de la voz sintética en condiciones ruidosas extremas.

3 AhoTTS Multilingüe

Para añadir nuevos idiomas a los ya disponibles se han integrado los módulos lingüísticos de código abierto diseñados por otras universidades. Aunque habría sido posible desarrollar un procesador lingüístico único para todos los idiomas esta posibilidad fue descartada debido a la dificultad de su implementación. Así, el procesador lingüístico del catalán se ha tomado de Festcat (Bonafonte et al., 2009), el del gallego de Cotovía (Rodríguez et al., 2012) y el del inglés de Festival (Taylor y Black, 1998). Para la síntesis, el motor que se emplea en todos los casos es el estadístico basado en HTS (Zen et, al., 2007). Debido a la manera peculiar en que cada uno de estos analizadores lingüísticos extrae la información y a la dependencia que de ella tienen las correspondientes voces sintéticas, ha sido necesario desarrollar nuevas voces para los idiomas recién incorporados.

Un diagrama general del sistema AhoTTS multilingüe diseñado puede verse en la Figura 2, en la que se observan los distintos módulos lingüísticos integrados para cada lengua.

El código fuente del sistema CTV multilingüe, junto con las voces disponibles, puede encontrarse en el siguiente repositorio de SourceForge http://sourceforge.net/projects/ahottsmultiling/.

3.1 Integración del Catalán

Para el idioma catalán se ha optado por el sistema Festcat. Festcat es el sistema CTV desarrollado por el grupo TALP de la UPC para la lengua catalana. Está desarrollado en base a Festival y sus diferentes módulos están escritos en el lenguaje de programación Lisp.

En AhoTTS multilingüe se emplean los módulos de procesado lingüístico de Festcat para obtener las etiquetas lingüísticas contextuales correspondientes al texto de entrada. Para poder emplear dichos módulos sin la necesidad de instalar previamente Festival, en el repositorio de AhoTTS multilingüe se suministran estas librerías ya compiladas para máquinas Linux de 64bits. El uso en otros sistemas operativos requerirá la compilación de las librerías adecuadas.

Para las llamadas a Festcat y evaluación de comandos Lisp dentro de AhoTTS se emplea la API C/C++ de la universidad de Edimburgo. Una vez obtenidas las etiquetas contextuales se

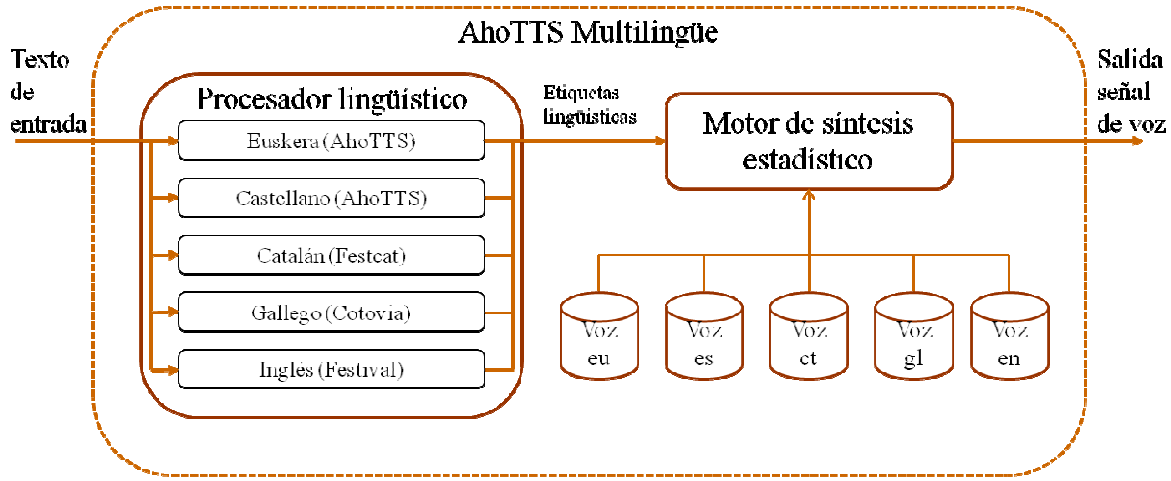

Figura 2: Diagrama general del AhoTTS Multilingüe

modifica el formato de las mismas para adecuarlo a la entrada del motor de síntesis. Después se introducen directamente en este módulo de igual manera que las generadas por los módulos lingüísticos propios de AhoTTS.

3.2 Integración del Inglés

En este caso se han incluido los módulos de procesado lingüístico de inglés proporcionados por la universidad de Edimburgo (Taylor y Black, 1998). De este modo se ha aprovechado el trabajo realizado para el catalán añadiendo fácilmente el inglés. Al igual que en el catalán es necesario adecuar el formato de las etiquetas lingüísticas al requerido por el motor de síntesis.

Es interesante destacar que cualquier módulo de procesado lingüístico desarrollado para Festival en otro idioma puede integrarse de manera sencilla en AhoTTS Multilingüe.

3.3 Integración del Gallego

El sistema escogido para la integración del gallego es Cotovia del grupo GTM de la UVIGO. Como el código del CTV está escrito en C/C++, simplemente se han integrado las funciones correspondientes al procesado lingüístico y generación de las correspondientes etiquetas en AhoTTS Multilingüe.

La salida que genera esté módulo sigue el formato ECESS (Pérez et al., 2006) basado en XML. Por tanto se ha adecuado dicha salida al formato de etiquetas de entrada del motor de síntesis.

3.4 Entrenamiento de Nuevas Voces

Previamente a la integración de los nuevos idiomas (el catalán, el gallego y el inglés) en AhoTTS, ya se disponía de modelos estadísticos para voces femenina y masculina en castellano y euskera generadas usando las bases de datos AhoSyn (Sainz et al., 2012b). También se disponía de una voz femenina para inglés entrenada a partir de la base de datos CMU ARCTIC (Kominek y Black, 2008). Por tanto, para el catalán y el gallego ha sido necesario desarrollar las correspondientes voces.

El entrenamiento de una nueva voz requiere un corpus fonéticamente balanceado y la transcripción de cada frase, así como las correspondientes grabaciones. A partir del audio se extraen los parámetros acústicos de la voz, y del texto se extraen las etiquetas lingüísticas correspondientes de forma que el sistema pueda aprender una relación estadística entre ambas. Las bases de datos para realizar dichos entrenamientos han sido cedidas por el grupo TALP de la UPC en el caso del catalán y por el grupo GTM de la UVIGO para el gallego.

Para entrenar voces para los nuevos idiomas se emplea HTS 2.2 (http://hts.sp.nitech.ac.jp/). Los datos principales sobre el tamaño de las bases de datos para la construcción de todas voces del sistema se resumen en la Tabla 1. En el caso de castellano y euskera el corpus para ambos géneros es el mismo en cada idioma.

Voz	Nº Frases	Nº Palabras	Duración aprox.
Voces castellano	3995	51380	6 horas
Voces euskera	3799	38544	6 horas
Inglés femenina	1132	10002	1 hora
Catalán femenina	3974	62314	6 horas
Catalán masculina	3692	58154	6 horas
Gallego masculina	1316	11235	1 hora

Tabla 1: Tamaño de las bases de datos usadas para la construcción de las voces

La parametrización acústica empleada consta de 39 coeficientes cepstrales en escala Mel junto con sus diferencias y sus segundas diferencias. También se extrae la frecuencia fundamental junto con su primera y segunda diferencia, así como un parámetro que indica el grado de sonoridad (más concretamente, la frecuencia máxima a la que la señal muestra armonicidad).

Como puede verse en la tabla 1, la cantidad de material disponible para el desarrollo de la voz gallega e inglesa es bastante inferior al disponible para las otras voces, a pesar de lo cual se ha considerado que es suficiente para obtener una calidad aceptable debido al método de síntesis empleado.

3.5 Transformación de Voces

Uno de los objetivos del proyecto en el que se enmarca el desarrollo de este sistema era la obtención de voces femeninas y masculinas

para todos los idiomas ya mencionados. Sin embargo, no fue posible obtener las bases de datos abiertas y libres de licencia en todos los casos. Por ello, y también por razones de economía de trabajo, se optó por aplicar técnicas de transformación de voces para completar el catálogo de las voces.

La transformación se ha llevado a cabo modificando los modelos estadísticos de la voz original a dos niveles: (i) se ha modificado el nivel medio de la frecuencia fundamental; (ii) se han aplicado técnicas de normalización del tracto vocal, lo que en el domino cepstral se traduce en un simple producto por una matriz como demuestran Pitz y Ney (2005). Para un valor adecuado de los parámetros de esta transformación, el resultado es una voz perceptualmente distinta a la original y que además mantiene un nivel de naturalidad comparable al de ésta.

La versión del sistema liberado dispone de las voces femeninas desarrolladas para los cinco idiomas.

3.6 API de desarrollo

En el repositorio de SourceForge junto con el código se proporciona además una API de desarrollo. Esta API permite incluir las funcionalidades básicas de AhoTTS multilingüe de manera sencilla en otros programas. También están incluidas aplicaciones de ejemplo para ilustrar el uso de esta API: un sistema autónomo y otro con arquitectura cliente/servidor.

De entre las características de las que dispone la API, las principales son:

- Permite cambiar la velocidad de lectura del texto en tiempo de síntesis.
- Proporciona las muestras de la salida para que el desarrollador las gestione de la manera que le convenga, ya sea guardándolas en un archivo de audio o enviándolas directamente a la tarjeta de sonido usando librerías del sistema operativo.
- Realiza el procesado frase a frase lo que permite el uso del sistema en aplicaciones en las que es necesario el procesado en tiempo real.

4 Conclusiones y Trabajos Futuros

Este artículo describe el sistema de conversión de texto a voz de código abierto desarrollado para las cuatro lenguas oficiales del estado más el inglés. El sistema del repositorio incluye además del código fuente del conversor multilingüe, las voces femeninas y una API de desarrollo para facilitar su integración en otras aplicaciones.

El hecho de que sea código abierto permite que cualquier persona interesada pueda descargarlo desde el repositorio donde se encuentra y utilizarlo para aprender, investigar o mejorarlo.

Este sistema multilingüe se ha desarrollado en el contexto del proyecto TV SOCIAL (http://tvsocial.ibercom.com/).

En el futuro se tiene pensado incluir como parte del procesado lingüístico un módulo previo de detección del idioma. De este modo se detectará automáticamente la lengua en la que está escrito el texto y se llamará directamente al módulo de procesado lingüístico correspondiente.

También se pretende crear una única voz multilingüe que incluya las particularidades fonéticas de todos los idiomas de manera que pueda usarse como voz única de todo el sistema.

5 Agradecimientos

Agradecemos al grupo TALP de la UPC y al grupo GTM de la UVIGO su ayuda y el material cedido para la creación de las voces catalanas y gallega respectivamente.

Queremos reconocer el trabajo de todas las personas que han colaborado en algún momento en el desarrollo de AhoTTS durante los últimos 20 años.

También agradecemos el trabajo realizado a todos los grupos que han liberado el código de sus sistemas CTV.

La migración del sistema a código abierto ha sido parcialmente financiada por el Gobierno Vasco (proyectos Ber2Tek, IE12-333 y Etorgai, ER-2010/00003), la empresa Eleka Ing. Ling. S.L. y por el Ministerio de Economía y Competitividad (Proyecto SpeechTech4All, TEC2012-38939-C03-03).

Bibliografía

Bonafonte, A., L. Aguilar, I. Esquerra, S. Oller, A. Moreno, 2009 "Recent Work on the FESTCAT Database for Speech Synthesis", Proc. SLTECH pp. 131-132.

Erro, D., I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, I. Hernáez, 2010, "HMM-based Speech Synthesis in Basque Language using HTS", Proc. FALA 2010 (VI Jornadas en Tecnología del Habla & II Iberian SLTech), pp. 67-70, (Vigo).

Erro, D., I. Sainz, E. Navas, I. Hernaez, 2011, "Improved HNM-based Vocoder for Statistical Synthesizers", Proc. Interspeech, pp. 1809-1812, (Florence).

Erro, D., T.C. Zorila, Y. Stylianou, E. Navas, I. Hernáez, 2013 "Statistical Synthesizer with Embedded Prosodic and Spectral Modifications to Generate Highly Intelligible Speech in Noise", Proc. Interspeech, (Lyon).

Hunt, A., A. Black, 1996 "Unit selection in a concatenative speech synthesis system using a large speech database", Proc. ICASSP, vol. 1, pp. 373–376.

Kominek, J., A Black, 2004 ”The CMU Arctic speech databases”, Proc. 5th ISCA Speech Synthesis Workshop, pp 223-224, Pittsburgh, PA.

Hernaez, I. 1995 “Conversión de texto a voz para el euskera basada en un sintetizador de formantes”, Tesis doctoral, UPV/EHU.

Ling, Z.H., L. Qin, H. Lu, Y. Gao, L.R. Dai, R.H. Wang, Y. Jiang, Z.W. Zhao, J.H. Yang, Y.J. Chen, G.P. Hu, 2007 "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007", Proc. Blizzard Challenge Workshop, Aug .

Navas, E., I. Hernáez, J. Sánchez, 2002 "Basque Intonation Modelling For Text To Speech Conversion", Proc. 7th International Conference on Spoken Language Processing (ICSLP), pp. 2409-2412, Denver.

Navas, E., I. Hernáez, J. Sánchez, 2002 "Modelo de duración para conversión de texto a voz en euskera", Procesamiento del Lenguaje Natural, vol. 29, pp. 147-152.

Navas, E , 2003 “Modelado prosódico del euskera batua para conversión de texto a habla”, Tesis doctoral, UPV/EHU.

Pérez, J., A. Bonaforte, H.U. Hain, E. Keller, S. Breueur, J. Tian, 2006 “ECESS Inter-Module Interface Specification for Speech Synthesis”, Proceedings of LREC Conference.

Pitz, M., H. Ney, 2005 “Vocal tract normalization equals linear transformation in cepstral space”, IEEE Trans. Speech and Audio Process., vol. 13(5), pp. 930-944.

Rodríguez, E., C. García, F. Méndez, M. Gozález, C. Magariños, 2012 “Cotovía: an Open Source Text-to-Speech System for Galician and Spanish”, Proc. Iberspeech 2012 (VII Jornadas en Tecnología del Habla & III Iberian SLTech), pp. 308-315, (Madrid).

Rodríguez, M.A., J.G. Escalada, D. Torre, 1998 ”Conversor multilingüe para castellano, catalán, gallego y euskera”, Procesamiento del lenguaje natural, Revista nº 23 pp19-23.

Sainz, I., D. Erro, E. Navas, J. Adell, A. Bonafonte, 2011 "BUCEADOR Hybrid TTS for Blizzard Challenge 2011", Proc. Blizzard Challenge Workshop, (Torino).

Sainz, I., D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, I. Odriozola, I. Luengo, 2010 "Aholab Speech Synthesizers for Albayzin2010", Proc. FALA 2010 (VI Jornadas en Tecnología del Habla & II Iberian SLTech), pp. 343-347, (Vigo).

Sainz, I., D. Erro, E. Navas, I. Hernáez, 2011 "A Hybrid TTS Approach for Prosody and Acoustic Modules", Proc. Interspeech, pp. 333-336.

Sainz, I., D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, , 2012a "Aholab Speech Synthesizer for Albayzin 2012 Speech Synthesis Evaluation", Proc. Iberspeech 2012 (VII Jornadas en Tecnología del Habla & III Iberian SLTech), pp. 645-652, (Madrid).

Sainz, I., D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga and I. Odriozola, 2012b “Versatile Speech Databases for High Quality Synthesis for Basque”, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), pp. 3308-3312.

Taylor, P., Black, A. and Caley, R, 1998 “The architecture of the Festival Speech Synthesis System”, Proc. 3rd ESCA Workshop on Speech Synthesis, pp. 147-151, Jenolan (Caves, Australia).

Zen, H., T Nose, J Yamagishi, S Sako, T Masuko, AW Black, K Tokuda, 2007 “The HMM-based speech synthesis system (HTS)

version 2.0", Proc. ISCA Workshop on Speech Synthesis (SSW6), pp. 294-299.

Zen, H., K. Tokuda, A. W. Black, 2009 "Statistical parametric speech synthesis", Speech Communication, Volume 51, Issue 11, pp. 1039-1064.