Procesamiento del Lenguaje Natural, num. 30 (2003), pp.125-126 recibido 1-03-2003; aceptado 20-05-2003

Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión Texto-Voz en Español

Autor: David Escudero
Universidad de Valladolid
Departamento de Informática
47011 Valladolid, España
descuder@infor.uva.es

Tutor: Valentín Cardeñoso
Departamento de Informática
Universidad de Valladolid
47011 Valladolid, España
valen@infor.uva.es

Resumen: En esta tesis se hace una propuesta de modelado de entonación original, que captura los patrones de entonación de un corpus y su variabilidad. La metodología de modelado se basa en representaciones paramétricas de la evolución del pitch en los grupos acentuales de un corpus y posterior modelado estadístico de los mismos. Se emplean técnicas de simulación para generar entonación sintética, reproduciendo la variabilidad observada en el corpus.
Palabras clave: Tecnologías del Habla, Conversión Texto Habla, Prosodia Computacional, Modelos de Entonación

Abstract: In this thesis we propose a technique for modelling intonation that aims to capture the paterns of intonation in a corpus and its variability. The methodology is based in parametric representations of the evolution of the pitch in the stress groups of a corpus and its statistical description. We use simulation techniques for the generation of synthetic intonation, showing the variability presented in the corpus.
Keywords: Speech Technology, Text to Speech, Computational Prosody, Intonation Modelling

1 Justificación y Objetivos

Ninguno de los sistemas actuales de conversión texto-voz (CTV) consigue generar voz con la misma naturalidad con la que lo hacemos los humanos. Esta falta de naturalidad se debe, esencialmente, a deficiencias de los modelos prosódicos empleados, generalmente excesivamente estáticos y simplistas. Esta tesis propone un conjunto de aportaciones al modelado estadístico de entonación y a la generación de habla sintética que tienen como objetivo común una mejora de la naturalidad de sistemas de CTV reales.

La variabilidad es una característica de la entonación que habitualmente se considera la dificultad esencial que hay que filtrar a la hora de modelar. Sin embargo, en esta tesis se sostiene la idea de que esta variabilidad es cualidad intrínseca de la entonación que debe ser explícitamente modelada, reproducida y aprovechada para corregir la monotonía característica de la mayoría de los sistemas de generación de habla sintética actuales.

El objetivo fundamental de este trabajo fue, por tanto, definir una metodología de modelado y análisis de la entonación basada en corpus que permita representar explícitamente la variabilidad real medida en ejemplares de locuciones correctas y la utilice con ventaja para generar voz sintética con un nivel de naturalidad superior al habitual. En lo que sigue se describe brevemente la metodología propuesta y el trabajo experimental realizado.

2 Metodología Propuesta

La propuesta de modelado de entonación desarrollada comprende cuatro bloques fundamentales: modelado a partir de corpus, representación cuantitativa basada en el grupo acentual como unidad de entonación, modelado estadístico y generación de entonación sintética basada en simulación de las observaciones.

En la figura 1 se muestra de forma esquemática la metodología de modelado y generación propuesta. Se segmenta el corpus disponible en grupos acentuales. Se elige el grupo acentual (GA) como unidad básica de entonación para el español porque está bien documentado que sobre él se proyectan los movimientos significativos de la entonación. Los perfiles de entonación de los GA se parametrizan empleando funciones de Bézier y los puntos de control de estas funciones

ISSN 1135-5948

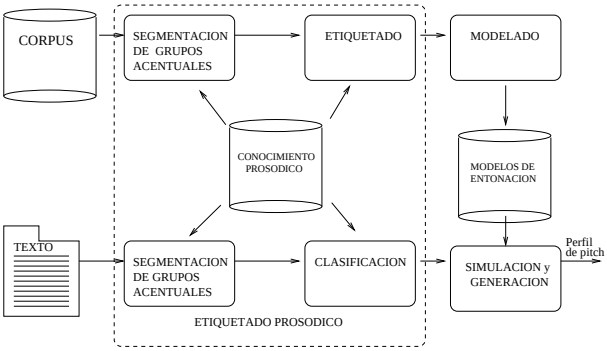

Figura 1: Esquema funcional de la metodología de modelado propuesta en esta tesis.

aproximantes al perfil de F0 se toman como parámetros de cada GA. Esto permite reflejar la evolución de la curva melódica con un número fijo de parámetros para cada GA con independencia de la duración y naturaleza del mismo. Como modelos de entonación adoptamos las distribuciones estadísticas de los parámetros de entonación en cada tipo de grupo acentual, así conseguimos reproducir adecuadamente las variabilidades presentes en el corpus que distinguen cada variante de unidad elemental de entonación considerada. Estos modelos son útiles para la etapa de generación de habla sintética, en tanto se disponga de un método de simulación para generar el conjunto de parámetros finales que ha de asociarse a cada uno de los GA presentes en el texto a sintetizar.

3 Investigación Realizada

Empleamos el corpus ESMA-UPC elaborado por el Grupo de Procesamiento de Señal de la Universidad Politécnica de Cataluña que contiene la cantidad y variedad de muestras adecuadas para proporcionar solidez a los resultados estadísticos. El corpus está segmentado en grupos acentuales y enriquecido con los atributos prosódicos habituales (tipo de acento, posición, ...).

Para evaluar la necesidad de considerar el entorno en que aparece cada grupo acentual y valorar el grado de aproximación de las curvas que resulta más adecuado a nuestros fines se analizaron diversas alternativas de aproximación de contornos de F0 y se diseñó un conjunto de métricas de comparación. Se seleccionó la aproximación de grado 3 a nivel de GA individual como la que mejor se ajusta a los objetivos, en comparación con las clásicas.

Se utilizaron clasificaciones de grupos acentuales inspiradas en propuestas recientes de otros autores para seleccionar el conjunto de rasgos prosódicos que mejor etiqueta cada GA y se valoraron empleando métricas objetivas de calidad que permitieron seleccionar la que mejor se adecúa al corpus de trabajo y a las limitaciones del método de generación de perfiles.

Se estudiaron dos alternativas de simulación basada en modelos para los parámetros de entonación: elección de valores medios de parámetros en las clases y superposición de variaciones aleatorias sobre la media, compatibles con los modelos generados a partir del corpus.

Se han desarrollado dos módulos software para generar perfiles de entonación que han sido incluidos en sendas arquitecturas de conversión texto-voz reales y se realizaron pruebas de evaluación objetiva y subjetiva de resultados.

4 Resultados y Conclusiones

Los resultados de las técnicas de evaluación aplicadas, y su uso en sistemas CTV reales, revelan que los modelos de entonación propuestos en nuestro trabajo sirven para generar perfiles de pitch de alta naturalidad. Las funciones de Bézier son un recurso de gran utilidad para parametrizar la entonación, con claras ventajas respecto a las técnicas de estilización clásicas. La validación de propuestas de clasificación de grupos acentuales y el modelado estadístico proporcionan instrumentos de enorme utilidad para extraer conocimiento prosódico a partir de un corpus, extremo en el que se sigue profundizando en etapas posteriores del trabajo. Las técnicas de simulación estudiadas permiten reproducir la variabilidad de los perfiles de pitch contenidos en un corpus, contribuyendo por tanto a mejorar la naturalidad en la generación de habla sintética. Los resultados de modelado de entonación y las técnicas de simulación de perfiles de F0 pueden ser incluidas en arquitecturas reales de conversión texto-voz con enorme facilidad, lo que pone de manifiesto la importante dimensión práctica de este trabajo de investigación.