DESCRIPCIÓN DEL CONVERSOR DE TEXTO A VOZ AHOTTS PRESENTADO A LA EVALUACIÓN ALBAYZIN TTS 2008

Iñaki Sainz, Inma Hernáez, Eva Navas, Jon Sanchez, Iker Luengo, Ibon Saratxaga, Igor Odriozola, Eneritz de Bilbao, Daniel Erro

Aholab Signal Processing Laboratory.
Departamento de Electrónica y Comunicaciones.
Universidad del País Vasco

RESUMEN

En el presente artículo se describen las características básicas del conversor de texto a voz (CTV) AhoTTS desarrollado por el grupo Aholab de la Universidad del País Vasco. AhoTTS es un sistema en el que tanto el módulo prosódico como el acústico están basados en técnicas por corpus. Asimismo se detalla el proceso de generación de una voz en castellano dentro de la campaña de evaluación Albayzin TTS 2008.

1. INTRODUCCIÓN

La campaña de evaluación Albayzin TTS 2008 tiene como propósito principal comparar las distintas técnicas e implementaciones de los sistemas participantes, partiendo de un base de datos común. Para ello, sigue la línea trazada por la evaluación internacional "Blizzard Challenge". La última edición de dicha evaluación se llevó a cabo tanto para el Inglés como para el Chino Mandarín, mientras que Albayzin TTS 2008 en esta su primera edición, se ha centrado únicamente en el Castellano.

Cada participante ha dispuesto de un periodo de 7 semanas para generar una voz a partir de la base de datos proporcionada por la UPC. Tras dicho periodo, se han sintetizado múltiples textos de test que serán evaluados de forma subjetiva bajo los siguientes criterios de calidad: Parecido con la voz original, Naturalidad e Inteligibilidad.

En este artículo se explican las características principales del CTV AhoTTS. En la sección 2 se describen los módulos que componen el sistema de síntesis. El proceso de generación de la voz se explica en la sección 3. Finalmente, se presentan unas conclusiones sobre todo el proceso en la sección 4.

2. DESCRIPCIÓN DEL SISTEMA

AhoTTS es el CTV que el grupo Aholab lleva desarrollando desde 1997. Implementado en C/C++ dispone de un arquitectura modular y multiplataforma. En la actualidad se han desarrollado voces en los siguientes idiomas: Euskera, Castellano e Inglés (para este último haciendo uso de los módulos de procesado lingüístico de Festival [1]). En las siguientes subsecciones se explicarán las características principales de cada uno de los módulos básicos que componen el sistema general: Procesado Lingüístico, Predicción Prosódica y Módulo Acústico.

2.1. Procesado Lingüístico

La función principal de este módulo es la de proporcionar una secuencia de fonemas a partir de un texto de entrada. Este proceso implica varias fases: normalización, delimitación de las frases, categorización, silabificado, acentuación y transcripción fonética. Aunque AhoTTS ya disponía de dichas funciones para el castellano, se han realizado mejoras o modificaciones en las tres últimas etapas.

2.2. Predicción de la Prosodia

Partiendo de la información del módulo precedente se pretende modelar la prosodia (entonación, duración y potencia) buscando imitar lo mejor posible la del locutor original. Por ello se han desarrollado modelos basados en el corpus proporcionado por la organización de Albayzin TTS 2008. Cabe destacar que no se ha desarrollado ningún modelo de inserción de pausas para la voz en castellano, por lo que únicamente se utiliza la puntuación ortográfica del texto de entrada.

2.2.1. Duración

La duración de cada fonema se predice mediante árboles CART para vocales, semivocales, consonantes sonoras y consonantes sordas. El entrenamiento se lleva a cabo en base a la siguientes características con una ventana que recoge los 2 fonemas anteriores y posteriores al actual: Fonema, vocal/consonante, vocales (altura, amplitud y redondez), consonantes

(sonoridad, clase, punto articulación…), acento, posición (sílaba, palabra, frase), etc.

2.2.2. Entonación

Se trata sin duda del rasgo prosódico más relevante en la calidad y naturalidad de la síntesis obtenida. AhoTTS dispone de tres modelos entonativos:

- *Modelo 1*: Una implementación muy simple de picos y valles.
- *Modelo 2*: Un modelo estadístico basado en árboles y curvas de Fujisaki que proporciona una entonación con una alta consistencia y naturalidad. Sólo está implementado para Euskera.
- *Modelo 3*: Modelo entonativo basado en corpus.

El sistema presentado en la evaluación Albayzin TTS 2008 hace uso del modelo 3, por lo que se procederá a detallar las características del mismo en las siguientes líneas.

Como todo modelo entonativo basado en la selección de unidades, para formar la curva final resultante extrae y concatena curvas de pitch naturales. A diferencia de la mayoría de sistemas, que utilizan como unidad básica el grupo acentual, se ha optado por una implementación similar a la de [2] en la que la unidad básica es el fonema.

- *Coste Objetivo*: A partir de la transcripción fonética de entrada y para cada fonema sonoro, se realiza una preselección de candidatos en base a las siguientes características: Fonema (coste nulo si el fonema es idéntico y en caso contrario ponderado por clases fonéticas), Sonoridad, Duración, Tipo de Grupo Acentual, Posición en el Grupo Acentual, Posición en la Sílaba, Tipo de Fonema Adyacente (vocal, semivocal, consonante sonora/sorda y pausa), Tipo de Grupo Fónico, Distancia al acento más cercano, Posición del Grupo Acentual dentro del Grupo Entonativo y Número de Grupos Acentuales dentro del Grupo Entonativo. Los pesos de los costes objetivo se entrenan siguiendo un esquema similar al propuesto en [3] usando regresión lineal múltiple. Para ello se define como distancia a predecir, la del contorno de pitch del fonema sonoro anterior, actual y siguiente, dando mayor peso a la forma del contorno que al valor absoluto del pitch.
- *Coste Concatenación*: Tras la fase de preselección, se computan los costes de concatenación para obtener la curva definitiva. Dichos costes incluyen: Distancia entre el 'siguiente contorno natural del fonema anterior' y el del actual y viceversa (distancia entre el 'contorno del fonema anterior' y el 'contorno natural anterior del fonema actual'), Diferencia de pitch entre los extremos cuando se trata de dos fonemas sonoros adyacentes, y Penalización por máximo salto de pitch entre sílabas adyacentes (calculado a partir de la media y desviación estandar de los saltos en la voz natural).

Finalmente, se interpola el pitch en los fonemas sordos, se suaviza entre fonemas sonoros adyacentes si fuera necesario, y se modifica ligeramente la duración de cada fonema sonoro interpolando la predicha por el modelo de duración correspondiente, con la duración del contorno entonativo seleccionado.

2.2.3. Energía

Dado que la potencia no se utiliza como coste objetivo durante la selección de unidades acústicas, se ha optado por no generar un modelo de energía para la voz en castellano.

2.3. Módulo Acústico

El módulo acústico combina las fases típicas de un sistema concatenativo basado en corpus: Preselección de Unidades, Programación Dinámica combinando los costes objetivo y de concatenación y la Síntesis de la secuencia de unidades seleccionadas para generar la onda de audio final.

2.3.1. Selección de Unidades

La unidad básica empleada por nuestro sistema es el semifonema, pero si existen suficientes candidatos (umbral situado en unos pocos centenares) hacemos uso de difonemas. De esta forma se establece un compromiso entre la flexibilidad prosódica que permite el hacer uso de semifonemas y la preservación de la naturalidad motivada por el uso de difonemas.

Haciendo uso del algoritmo de Viterbi se busca la secuencia de unidades del corpus que minimice la función coste compuesta por subcostes objetivo y de concatenación tal y como se muestra en las siguientes fórmulas:

$$C(t_1...t_n,u_1...u_n)=\alpha\sum_{i=1}^{n}C^T(t_i,u_i)+(1-\alpha)\sum_{i=1}^{n-1}C^C(u_i,u_{i+1})$$

$$C^T(t_i,u_i)=\sum_{j=0}^{P}w_j^T C_j^T(t_i,u_i)$$

$$C^C(u_i,u_{i+1})=\sum_{j=0}^{Q}w_j^C C_j^C(u_i,u_{i+1})$$

Donde t_i identifica las unidades objetivo y u_i las candidatas. C^T y C^C representan los costes objetivo y concatenación respectivamente; w_j es el j-ésimo peso que pondera una de las subfunciones existentes: P subfunciones de coste objetivo y Q de concatenación.

El coste objetivo está formado por la suma ponderada de los siguientes subcostes aplicados a nivel de semifonema y normalizados entre 0 y 1 (coste máximo):

- *Trifonema*: Valor discreto para potenciar el uso de unidades consecutivas en el corpus.
- *Contexto*: En una ventana de 5 fonemas, conjunto de valores discretos que caracterizan los tipos de fonemas adyacentes.
- *Pitch*: Distancia euclídea del contorno entonativo normalizando la duración.
- *Duración*: Valor absoluto de la diferencia de longitud. Se tiene en cuenta la posibilidad de modificar ligeramente la duración de unidades sonoras durante la generación de la forma de onda.
- *Acento*: Distancia a la sílaba acentuada más próxima.
- *Tipo de grupo fónico*: Interrogativo, inacabado, exclamativo, enunciativo, etc. Se ponderan especialmente las unidades finales y también la iniciales en las oraciones interrogativas.
- *Posición*: Posición relativa de la unidad dentro del grupo fónico.
- *Posición en la palabra*: Las unidades se agrupan en 4 categorías (inicio, medio, final y única) a nivel tanto de palabra como de sílaba.
- *Sonoridad*: Penaliza unidades sonoras marcadas como sordas.
- *Calidad fonética*: Penaliza unidades que si bien no son marcadas como "fuera de rango" su distancia acústica es superior a un umbral respecto al centro del cluster.

Los pesos de los costes objetivo se ajustan de forma automática utilizando un método similar al utilizado en el módulo prosódico. Se mide la distancia acústica entre unidades del corpus para tratar de predecirla como la suma ponderada de las funciones coste; resolviendo el valor de los pesos como un problema de regresión lineal múltiple.

Tras realizar una preselección con las unidades de menor coste objetivo, se calculan los costes de concatenación entre unidades no consecutivas en el corpus en base a los siguientes criterios:

- *Pitch*: Diferencia de pitch en el punto de concatenación.
- *Rango de pitch*: Para controlar saltos excesivos de pitch entre sílabas adyacentes y normalizando el coste respecto a los valores medios medidos para la voz original.
- *Duración*: Calculada únicamente a nivel de fonema (sumando duraciones de semifonema izquierdo y derecho) como diferencia respecto a la predicha por el modelo de duración.
- *Potencia*: Potencia en los extremos a concatenar y potencia media para unidades sonoras a nivel de fonema.
- *Sonoridad*: Penaliza unión entre unidades sonoras marcadas como sordas que no sean consecutivas. Para evitar ruidos de concatenación debidos a marcas de pitch erróneas.
- *Punto de unión*: Se penaliza ligeramente las uniones en partes no estacionarias, es decir, transiciones entre fonemas.
- *Distancia acústica*: Distancia euclídea entre la última y primera OLA de las unidades a concatenar. Se parametriza mediante 13 coeficientes MFCC añadiendo primeras y segundas diferencias. Para normalizar los valores se computan previamente las distancias medias de las transiciones entre semifonemas de la voz original.

Los pesos relativos a los costes de concatenación son ajustados de forma manual, aunque no se modificaron los valores existentes para las voces en Euskera, salvo el coste α que pondera la importancia entre los costes objetivo y de concatenación.

2.3.2. Generación de la forma de onda

Para mantener al máximo la naturalidad de las unidades seleccionadas, no se realiza ningún tipo de modificación de pitch, suavizando únicamente las uniones con información del cierre del pulso glotal. Sí que se realiza en cambio, una modificación de la duración de las unidades sonoras cuando la diferencia respecto al objetivo excede un umbral. Así como una ligera modificación de la energía para evitar cambios bruscos de volumen.

3. CONSTRUCCIÓN DE LA VOZ

Para la generación de la voz se partía del corpus upc_esma [4] grabado en castellano por una locutora. El corpus de 1 hora y 45 minutos de duración, está formado por 3 tipos de textos: frases fonéticamente balanceadas (30 minutos), párrafos fonéticamente balanceados (30 minutos) y párrafos literarios con una mayor variación prosódica (45 minutos).

Junto a los ficheros de audio se ha proporcionado la segmentación fonética, marcas a periodo de pitch, y señal del laringógrafo. Para la construcción de la voz sólo se ha utilizado la información relativa a la segmentación fonética, añadiendo de forma automática los alófonos aproximantes (B,D,G) y alguna otra diferencia respecto a nuestro transcriptor para castellano.

Debido a restricciones temporales no se llevó a cabo ningún tipo de revisión manual de las transcripciones. En cambio se realizó una detección automática de outliers en base a la siguiente información: *Score* devuelto por el segmentador, Duraciones, y Distancia Acústica de las unidades respecto al centro del cluster para cada fonema. Dicha información es utilizada también en el coste objetivo "calidad fonética", para evitar en la medida de lo posible, errores de etiquetado y/o pronunciaciones pobres.

El resto del procesado está igualmente automatizado y comienza con la normalización del

audio en base a la potencia media de las vocales. La curva entonativa es estimada mediante un sistema propio [5] basado en información cepstral y programación dinámica, que posteriormente se estiliza para cada semifonema mediante 3 puntos (inicio, punto más significativo y final). Para el marcado a periodo de pitch se utiliza la herramienta *epochs* de la suite *ESPS* corrigiendo en una etapa posterior, errores de pitch "halving/doubling" mediante la comparación del pitch local con nuestra estimación del contorno entonativo. Con la ayuda de la aplicación *sig2fv* del paquete *speech tools* se obtienen los coeficientes MFCC utilizados tanto en el coste de concatenación, como para entrenar los pesos de los costes objetivo y detectar outliers.

El resto de información necesaria se extrae a partir del módulo de procesamiento lingüístico.

4. CONCLUSIONES

En el presente artículo se han descrito las características esenciales del sistema AhoTTS, así como el proceso necesario para generar una nueva voz. Ésta ha sido la primera voz en castellano desarrollada en el laboratorio, y aunque a priori los resultados han sido bastante satisfactorios, existe un amplio margen de mejora.

Si bien todos los módulos del sistema pueden ser objeto de dicha mejora, quizá la generación de onda sea nuestro "talón de Aquiles". La solución pasa por la utilización de algún tipo de codificación que permita realizar modificaciones prosódicas y suavizado espectral en las concatenaciones, con poca degradación de la naturalidad.

5. AGRADECIMIENTOS

Agradecer el esfuerzo realizado por todos los sistemas participantes inscritos en la campaña de evaluación Albayzin TTS 2008.

6. BIBLIOGRAFÍA

[1] Taylor, P., Black, A. and Caley, R, "The architecture of the Festival Speech Synthesis System", *3rd ESCA Workshop on Speech Synthesis*, pp. 147-151, Jenolan Caves, Australia, 1998.

[2] Raux, A., Black, A., "A unit selection approach to f0 modeling and its application to emphasis", *Proc. of ASRU 2003*, St Thomas, US Virgin Is, 2003.

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

[4] Bonafonte, A., Moreno, A., "Documentation of the upc_esma spanish database", *TALP Research Center*, Universitat Politecnica de Catalunya, Carcelona; Spain , 2008.

[5] Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., Sainz, I., "Evaluation Of Pitch Detection Algorithms Under Real Conditions", *Proc. of 32nd IEEE ICASSP*, pp. 1057-1060, Honolulu, 2007.