ADAPTACIÓN DEL CTH-URL PARA LA COMPETICIÓN ALBAYZIN 2008

Carlos Monzo, Lluís Formiga, Jordi Adell, Ignasi Iriondo, Francesc Alías y Joan Claudi Socoró

GPMM - Grup de Recerca en Processament Multimodal
Enginyeria i Arquitectura La Salle - Universitat Ramon Llull
C/ Quatre Camins 2, 08022 Barcelona, Spain

RESUMEN

En esta comunicación describimos el sistema de síntesis de voz presentado a la competición Albayzin 2008. Es un sistema que sigue un esquema clásico de concatenación de unidades basado en corpus. Cabe destacar que los costes de selección se han ajustado mediante un método basado en algoritmos genéticos y que no se ha utilizado ningún sistema de predicción prosódica. Se construyeron dos sistemas preliminares que diferían en el algoritmo de generación de forma de onda escogiendo el que se presenta a la competición mediante un test perceptual.

1. INTRODUCCIÓN

La investigación sobre síntesis de voz en la Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle se inició en los ochenta con trabajos sobre síntesis articulatoria y por formantes [1, 2, 3].

Más tarde se optó por la síntesis concatenativa basada en difonemas. Se implementó un sintetizador en catalán basado en la técnica *TD-PSOLA* [4, 5], que ha sido la base para los sistemas desarrollados posteriormente hasta la actualidad. Desde entonces se ha mejorado la selección de textos a ser usados durante el proceso de grabación, la creación de reglas para la transcripción fonética (especialmente para castellano), la segmentación de las unidades y el marcado de *pitch*. Por otro lado, se han realizado investigaciones en el campo del modelado prosódico[6], ajuste de pesos para la función de selección [7, 8] y nuevas parametrizaciones basadas en cualidad de la voz [9].

La investigación del grupo los últimos años se ha basado en disminuir el coste de producción, segmentación y puesta a punto de los corpus de voz desarrollando nuevas herramientas de etiquetado automático. Dichas técnicas de rápida puesta a punto han sido desarrolladas dentro del proyecto europeo SALERO (*Semantic AudiovisuaL Entertainment Reusable Objects*). Dicho proyecto trata de conseguir un flujo producción para juegos, películas y televisión en diferentes medios de manera rápida, cualitativa y económica mediante la combinación de gráficos por ordenador, tecnologías de lenguaje y síntesis, tecnologías de web semántica así como búsqueda y recuperación basada en contenido.

El sistema presentado en esta evaluación está compuesto por tres módulos: el de transcripción fonética, selección de unidades y generación de forma de onda. La transcripción fonética se realiza mediante reglas. La selección de unidades se lleva a cabo mediante un algoritmo de programación dinámica y la generación de forma de onda mediante concatenación y modificación (en algunos casos) de la señal. La versión presentada en esta competición prescinde excepcionalmente del módulo de estimación prosódica [6] y considera la prosodia inherente en el corpus de voz de las unidades seleccionadas parcialmente siguiendo lo descrito en [10]. De esta forma el algoritmo de selección debe asegurar que la prosodia generada sea natural. Para ello debe ser capaz de escoger las unidades de forma que un elevado porcentaje sean seleccionadas consecutivamente, lo que conlleva aprovechar la variabilidad prosódica implícita en el corpus.

En las secciones 2, 4 y 5 se describen los módulos del sistema de síntesis. Además, en la sección 3 se explica el tratamiento del corpus para la generación del inventario de unidades y en la sección 6 el procedimiento seguido para selección del sistema definitivo.

Este trabajo ha sido subvencionado parcialmente por el proyecto SALERO (IST FP6-027122) de la Comisión Europea.

2. TRANSCRIPCIÓN FONÉTICA

El *phone-set* utilizado en nuestro sintetizador deriva de SAMPA [11]. El módulo de transcripción fonética consiste en un motor de reglas [12]. Las reglas actúan sobre una estructura de datos, que consiste en una lista de los pares grafema-fonema en un enunciado. Es posible utilizar una regla de inserción (I) o de borrado (D). La reglas se aplican sólo cuando la evaluación (E) de una cualidad de un fonema da un resultado positivo.

$$E(gr ==' h') \rightarrow D(gr) \quad (1)$$
$$E(gr ==' x') \rightarrow I(/ks/) \quad (2)$$

La regla (1) indica que se debe borrar el grafema $(gr)\,'h'$ y la (2) que el grafema $'x'$ se transforma en los fonemas $/ks/$ por lo que se produce una inserción. Para excepciones el sistema incluye un diccionario que se consulta antes de aplicar el motor de reglas.

3. CREACIÓN DEL INVENTARIO DE UNIDADES

3.1. Segmentación y Etiquetado

3.1.1. Segmentación por fonemas más detección de silencios

La segmentación es el proceso por el cual se etiqueta el corpus de voz indicando los límites temporales a nivel de fonema. Dentro del grupo de investigación se ha evolucionado el proceso de segmentación, incorporando mejoras respecto del primer marcador realizado [13], tanto en lo que se refiere a la calidad de marcado como en la facilidad de uso gracias a la creación de interfaces de usuario e independizándolo de la lengua de interés. Actualmente, el entrenamiento necesario y el posterior proceso de etiquetado está basados en el uso de modelos ocultos de *Markov* (HMM). Para ello se dispone de código propietario desarrollado en *Matlab®* que hace uso a su vez de la herramienta HTK (*Hidden Markov Model Toolkit*) [14].

El corpus que se ha suministrado es un corpus locutado por una voz femenina, de estilo neutro que consta de 776 archivos

de alrededor 6000 palabras y En lo que se refiere al análisis del corpus subministrado para la competición, esencialmente se ha realizado el control de la aparición y omisión de silencios respecto a los indicados en cada uno de los enunciados, independizando así el hecho que el locutor realizara las pausas de modo acorde al que los textos exigían. Por otro lado, los sonidos oclusivos se tratan de forma especial, de tal manera que el golpe de voz (*burst*) y el silencio previo se modelan como unidades diferentes.

3.1.2. Marcado de pitch

Se ha realizado en dos fases. En la primera se sitúan las marcas sobre zonas sonoras, utilizando para ello un marcador basado en el algoritmo RAPT [15]. En segundo lugar se aplica un postprocesado sobre las marcas llamado *Pitch Marks Filtering Algorithm* (PMFA) [16], con el que se consigue un marcado fiable, minimizando a su vez la existencia de inserciones y omisiones. Finalmente, se obtiene como resultado el marcado de zonas sordas y sonoras sin transiciones bruscas.

3.2. *Prunning* de la base de datos

Una vez creada la base de datos, con la información de duraciones (segmentación) y de F_0 media (*pitch*), se realiza un análisis estadístico para cada una de las unidades de forma que se descarten, durante la síntesis, aquellos casos que sean considerados erróneos (p.ej. *outliers*). El criterio considerado se basa en el hecho que todos aquellos valores, de duración y F_0, que estén fuera del margen ± 1.5 veces el valor de mediana calculado sobre todo el corpus para cada unidad serán descartados. A partir de este procedimiento, por tanto, se dispone de una lista (*black list*) donde se identifica la unidad que no se desea utilizar durante la síntesis, el archivo a la que pertenece y su posición dentro del enunciado.

En cuanto a los resultados obtenidos sobre las duraciones y F_0 medio se obtuvieron los siguiente porcentajes: 2.71% de valores descartados respecto a duración media y 1.10% respecto a la F_0 media.

4. OPTIMIZACIÓN DE COSTES DE SELECCIÓN

La selección de unidades se realiza mediante un algoritmo de programación dinámica que minimiza una función de coste para una serie de i unidades seleccionadas del corpus [17]. Al no haber predicción prosódica se omiten los costes de *target* y solo se tienen en cuenta los costes de concatenación. Según se puede observar en la ecuación (3), la matriz de costes es ponderada por un vector de pesos que intenta correlar los costes con la calidad final de la señal.

$$Csel_{\{U_1,U_2,\cdots,U_N\}} = \sum_{i=1}^{N-1}\sum_{j=1}^{3}\omega_j SC_j(i,i+1) \qquad (3)$$

En la ecuación (3) se presenta un ejemplo de cálculo de la función de coste, donde SC_{ij} corresponde al subcoste de seleccionar la unidad i según la parametrización $j = \{$PIT C, ENE C, MFCC C$\}$ y ω_j a su peso correspondiente.

4.1. Costes de concatenación y su normalización

Los costes de concatenación con los que se trabaja son:

- PIT C: Subcoste de *pitch* de concatenación. Determina la diferencia de F0 en el punto de concatenación de la unidad.
- ENE C: Subcoste de energía de concatenación. Es la diferencia de nivel energético de las unidades a concatenar en el punto de concatenación mencionado anteriormente.
- MFCC C: subcoste espectral de concatenación. Su cálculo se basa en la estimación del espectro mediante su parametrización cepstral en la escala Mel (en inglés, *Mel Frequency Cepstral Coefficients (MFCC)*. Se utilizan 24 coeficientes cepstrales más sus derivadas calculadas sobre una ventana de 20ms en el punto de concatenación.

La normalización de los costes para que sean comparables se basa en la aplicación de una función sigmoidea, debido a que la cantidad de valores atípicos no permiten una normalización MAX-MIN (la normalización sigmoidea pondera la parte lineal central de la distribución de valores). En las ecuaciones (4) y (5) se detalla la normalización siendo P_{ij}^R y $P_{(i+1)j}^L$ los subcostes de las unidades i e $i+1$ según la parametrización j y en su última (R - *right*) y primera (L - *left*) tramas, respectivamente. Adicionalmente $SC_j(u_i, u_{i+1})$ representa el coste de concatenación j de las unidades i e $i+1$ una vez normalizado. Dicha normalización se basa en el cálculo de las diferencias de los parámetros normalizándolas respecto a la desviación del subcoste (σ_{X^c}) sobre una función sigmoidea [18].

$$X^c(u_i,u_{i+1}) = \sum_{1}^{N}|P_{ij}^R - P_{(i+1)j}^L| \qquad (4)$$

$$SC_j(u_i,u_{i+1}) = 1 - e^{-\left(\frac{X^c(u_i,u_{i+1})}{\sigma_{X^c}}\right)^2} \qquad (5)$$

Asimismo, las estadísticas en el proceso de normalización se obtienen para cada una de las unidades analizadas, y se aplican evitando el problema del impacto de sesgar los subcostes debido a una normalización global de los mismos considerando todas las unidades del corpus.

Los pesos utilizados se han obtenido mediante una técnica de regresión que utiliza algoritmos genéticos por torneo (tGA) [19]. En este caso, se utilizaron los pesos normalizados que se calcularon para un corpus propio en catalán, de 20 minutos de duración y 1.207 unidades.

5. GENERACIÓN DE FORMA DE ONDA

Para la competición Albayzín 2008 se pusieron a punto dos sistemas de generación de forma de onda para la arquitectura del CTH explicada anteriormente. El primer sistema implementaba una síntesis basada en *Overlap and Add* en el dominio temporal (*TD-PSOLA*) [20] mientras que el segundo simplemente realizaba concatenación directa de la señal según sus marcas de *pitch* (*RAW*).

La modificación de la señal mediante *TD-PSOLA* se ha aplicado solamente a las unidades sonoras ralentizando su velocidad un 10% (aumentando su duración). Para alargar unidades se hace una interpolación de las tramas.

La concatenación *RAW* no hace ningún tipo de ventaneo y se concatena la señal en el paso por cero más cercano a la marca de *pitch*. La ventaja de este tipo de concatenación se basa en mantener la propia naturalidad de la señal. Dicho sistema, asume que ante una selección de unidades óptima, no es necesaria ninguna normalización o modificación de la señal y así se puede aprovechar la velocidad, volumen y entonación de la grabación original [21].

6. SELECCIÓN DEL SISTEMA PRESENTADO

Debido a la restricción de presentar un sólo sistema por equipo participante, se hizo una selección del sistema de síntesis más aceptado a nivel perceptual. A este efecto se escogieron al azar cinco frases de las 350 de test que se tenían que generar, se sintetizaron con *TD-PSOLA* y con *RAW* para que el usuario final escogiera entre ellas (según la naturalidad y inteligibilidad conseguidas).

15 usuarios, tanto expertos como no expertos en tecnologías del habla, realizaron la prueba cuyos resultados se pueden observar en la figura 1.

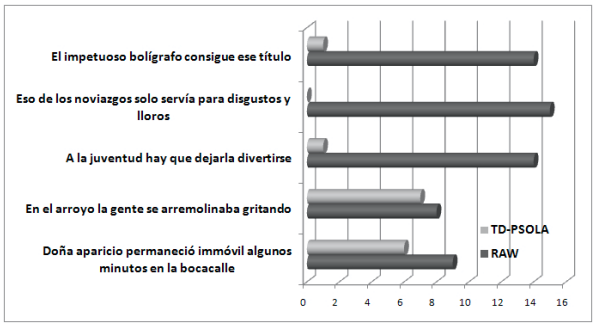

Figura 1. Porcentaje de preferencia del test perceptual para las cinco frases escogidas del test.

Analizando los resultados se puede observar que el número de votaciones entre *RAW* y *TD-PSOLA* es muy parecido en dos de las cinco frases donde siempre acaba ganando por la mínima *RAW*. En las otras frases *RAW* peresenta una abrumadora mayoría. Por este motivo, se decidió presentar a la evaluación el sistema *RAW*.

7. CONCLUSIONES

Desde el punto de vista de la creación de la voz se puede concluir que el tiempo de puesta a punto ha sido bastante rápido y eficiente. Para cuantificar el tiempo destinado para ello, serían en torno a 20 horas de parte de un técnico y las necesarias, en función de la longitud del corpus y de la potencia de cálculo de la máquina, para su parametrización. Cabe decir que el tiempo y coste destinado ha sido el más óptimo según nuestra experiencia, ya que para cada nueva voz desarrollada las herramientas para su etiquetado han sido mejoradas y los procesos más automatizados según el proyecto SALERO detallado en la introducción.

Cabe deducir, debido al resultado de las pruebas perceptuales, que al no disponer de modelo prosódico, la técnica de *TD-PSOLA* no mejora, y en algunos casos empeora, la calidad perceptual de la señal.

Es la segunda voz en castellano utilizada para un dominio genérico con nuestro conversor texto-habla. Nuestro primer conversor texto-habla utilizó una voz expresiva con distintas emociones el cual ha dado lugar a diferentes publicaciones y proyectos [22, 6, 23].

8. TRABAJO FUTURO

La estrategia de futuro de nuestro CTH pasa por incorporar técnicas de modificación de la señal basadas en modelos harmónicos más ruido, mejorar el procesado de la señal según *TD-PSOLA* teniendo en cuenta la especifidad de cada unidad, incorporar costes lingüísticos en coste de *target*, incorporar el trabajo prototipo de selección de unidades basado en medidas perceptuales detallado en [7, 8] y incorporar el modelo prosódico expresivo detallado en [6].

9. BIBLIOGRAFÍA

[1] Josep Martí, *Estudi acústic del català i síntesi automàtica per ordinador*, Ph.D. thesis, Universidad de Valencia, Valencia, España, 1985.

[2] Josep Martí, *Reconocimiento automático del habla*, chapter Síntesis del habla: Evolución histórica y situación actual, Boixareu Marcombo, 1987.

[3] Josep Martí, "Estado actual de la síntesis de voz," in *Estudios de Fonética Experimental*, 1990, vol. 4, pp. 147–168.

[4] Joan Camps, Gerard Bailly, y Josep Martí, "Synthèse à partir du texte pour le catalan," in *Proc. 19èmes Journées d'Études sur la Parole*, Bruselas, Francia, 1992, pp. 329–333.

[5] Roger Guaus, Francesc Gudayol, y Josep Martí, "Conversión textovoz mediante síntesis PSOLA," in *Jornadas Nacionales de Acústica*, Barcelona, España, 1996, pp. 355–358.

[6] Ignasi Iriondo, Joan Claudi Socoró, y Francesc Alías, "Prosody Modelling of Spanish for Expressive Speech Synthesis," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, EUA, Abril 2007, vol. 4, pp. 821–824.

[7] Lluís Formiga y Francesc Alías, "Extracting User Preferences by GTM for aiGA Weight Tuning in Unit Selection Text-to-Speech Synthesis," in *Computational and Ambiental Intelligence - Proceedings on 9th International Work-Conference on Artificial Neural Networks (IWANN)*, 2007.

[8] Francesc Alías, Xavier Llorà, Lluís Formiga, Kumara Sastry, y David E. Goldberg, "Efficient interactive weight tuning for TTS synthesis: reducing user fatigue by improving user consistency," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, Francia, 2006, vol. I, pp. 865–868.

[9] Carlos Monzo, Francesc Alías, Ignasi Iriondo, Xavier Gonzalvo, y Santiago Planet, "Discriminating Expressive Speech Styles by Voice Quality Parameterization," in *Proc. of the 16th International Congress of Phonetic Sciences(ICPhS)*, Saarbrücken, Alemania, Abril 2007, pp. 2081–2084.

[10] Francesc Alías, Ignasi Iriondo, Lluís Formiga, Xavier Gonzalvo, Carlos Monzo, y Xavier Sevillano, "High quality Spanish restricted-domain TTS oriented to a weather forecast application," in *Proc. of the 9th International Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal, 2005, pp. 2573–2576.

[11] John C. Wells, *SAMPA computer readable phonetic alphabetHandbook of Standards and Resources for Spoken Language Systems*, chapter SAMPA computer readable phonetic alphabet, pp. Part IV, section B, Berlin and New York: Mouton de Gruyter, 1997.

[12] Ignasi Iriondo, *Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva*, Ph.D. thesis, Universitat Ramón Llull, 2008.

[13] Francesc Alías y Ignasi Iriondo, "Segmentador de fonemas en catalán basado en DHMM," in *Actas del 16th Simposium Nacional de la Unión Científica (URSI)*, Madrid, España, 2001, pp. 149–150.

[14] "HTK," in *Recuperado el 19 de 09 de 2008, de http://htk.eng.cam.ac.uk*, 2008, pp. 149–150.

[15] David Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *W. B. Kleijn y K. K. Paliwal (eds.). Speech Coding and Synthesis. Amsterdam, NL: Elsevier Science*, 1995, pp. 495–518.

[16] Francesc Alías, Carlos Monzo, y Joan C. Socoró, "A Pitch Marks Filtering Algorithm based on Restricted Dynamic Programming," in *Proc. of International Conference on Speech and Language Processing (ICSLP)*.

[17] Andrew Hunt y Alan W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, EUA, 1996, vol. 1, pp. 373–376.

[18] Albert Febrer, *Síntesi de la parla per concatenació basada en la selecció*, Ph.D. thesis, Universitat Politècnica de Catalunya, Gener 2001.

[19] Francesc Alías y Xavier Llorà, "Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis," in *Proc. of the 8th European Conference on Speech Communication and Technology (EuroSpeech)*, Geneve, Suiza, 2003, pp. 1333–1336.

[20] Eric Moulines y Francis Charpentier, "Pitch-Sinchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Speech Communication*, 1990, vol. 9, pp. 453–467.

[21] Marcello Balestri, Alberto Pacchiotti, Silvia Quazza, Pier Luigi Salza, y Stefano Sandri, "Choose the best to modify the least: a new generation concatenative synthesis system," in *Proc. of the 6th European Conference on Speech Communication and Technology (EuroSpeech)*, 1999, pp. 2291–2294.

[22] Luigi Ceccaroni, Paloma Martínez, Josefa Z. Hernández, y Xavier Verdaguer, "IntegraTV-4all: an interactive television for all," in *Proc. of 1st International Symposium on Ubiquitous Computing and Ambient Intelligence (UCAmI'05)*, 2005.

[23] Francesc Alías, Xavier Sevillano, Joan Claudi Socoró, y Xavier Gonzalvo, "Towards high quality next-generation Text-to-Speech synthesis: a Multidomain approach by automatic domain classification," *IEEE Transactions on Audio, Speech and Language Processing (Special issue on New Approaches to Statistical Speech and Text Processing)*, vol. 16 (7), pp. 1340–1354, 2008.