Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 67-74 recibido 14-02-08, aceptado 03-03-08

Predicción estadística de las discontinuidades espectrales del habla para síntesis concatenativa

Statistical prediction of spectral discontinuities of speech in concatenative synthesis

Manuel Pablo Triviño y Francesc Alías
GTAM – Grup de Recerca en Tecnologies Audiovisuals i Multimèdia
Enginyeria i Arquitectura La Salle. Universitat Ramon Llull
Quatre Camins, 2. 08022 Barcelona, España
{st08726, falias}@salle.url.edu

Resumen: La estimación de discontinuidades espectrales es uno de los mayores problemas en el ámbito de la síntesis concatenativa del habla. Este artículo presenta una metodología basada en el estudio del comportamiento estadístico de medidas objetivas sobre uniones naturales. El objetivo es definir un proceso automático para seleccionar qué medidas emplear como coste de unión para sintetizar un habla lo más natural posible. El artículo presenta los resultados objetivos y subjetivos que permiten validar la propuesta.
Palabras clave: Medida objetiva, discontinuidad espectral, tipificación, correlación

Abstract: The estimation of spectral discontinuities is one of the most common problems in speech concatenative synthesis. This paper introduces a methodology based on analyzing the statistical behaviour of objective measures for natural concatenations. The main goal is defining an automatic process capable of including the most appropriate measures as concatenation cost to generate high quality synthetic speech. This paper describes both the objective and subjective results for validating the proposal.
Keywords: Objective measure, spectral discontinuity, standardization, correlation.

1 Introducción

Este trabajo se ubica en el ámbito de la generación de habla sintética a partir de texto o conversión de texto en habla (CTH). Existen distintas técnicas para obtener voz a partir de un texto cualquiera. Una de ellas es la síntesis por concatenación de unidades, en la que el habla sintetizada se genera uniendo segmentos de voz previamente grabados en un corpus. Uno de los problemas inherentes de este tipo de síntesis concatenativa es la aparición de discontinuidades audibles que se producen al unir las unidades acústicas (fonemas, difonemas, etc.).

En este contexto, la CTH basada en selección de unidades trabaja con corpus de voz de dimensión considerable (mayor a 1 hora de voz) (Hunt, 1996). Como su nombre indica, esta técnica se basa en seleccionar los segmentos del corpus que permitan generar un habla sintetizada lo más natural posible. El proceso de selección considera la bondad de la unión de las unidades a seleccionar para minimizar la presencia de discontinuidades en el habla sintética mediante criterios de coste basados en medidas objetivas (Hunt, 1996). La bondad de estas medidas vendrá determinada por su capacidad para detectar discontinuidades espectrales perceptibles.

Hasta el momento, la dificultad que conlleva mapear esta subjetividad provoca que todavía no se haya definido una medida objetiva única capaz de estimar el grado audible de una discontinuidad producida al concatenar dos unidades acústicas cualesquiera. Por ello, en la literatura sobre el tema se pueden encontrar diversos estudios que presentan resultados divergentes. En (Wouters, 1998) se concluye que la mejor distancia es la Euclídea aplicada sobre coeficientes MFCC (o incorporando sus derivadas). Sin embargo, en (Klabbers, 2001) se argumenta que la mejor predicción se consigue con la combinación de la distancia de Kullback-Leibler y los coeficientes LPC, mientras que en (Stylianou, 2001) se apuesta por la misma distancia pero con coeficientes FFT. Por su parte, en (Donovan, 2001) se

ISSN 1135-5948

define una medida basada en la distancia de Mahalanobis que mejora los resultados obtenidos en la literatura hasta el momento. Posteriormente, en (Vepa, 2006) el mejor resultado se obtiene para un coste basado en coeficientes LSF (*Line Spectral Frequencies*), propuesta que se completa con un método de interpolación lineal de concatenación de unidades usando también LSFs.

Desde otro punto de vista, se pueden encontrar trabajos que, además de estudiar las medidas objetivas, incorporan métodos de clasificación o regresión de las unidades acústicas. En (Syrdal, 2005) se aplica regresión lineal y CART (*Classification and Regression Trees*) a partir del etiquetado fonético y espectral del corpus. Se concluye que la agrupación por variables fonéticas permite una mejor predicción de las discontinuidades.

Ante la dificultad de detectar las discontinuidades a través de medidas objetivas de forma fidedigna, últimamente han aparecido nuevas propuestas con un enfoque distinto: el empleo de modelos harmónicos y componentes AM-FM (Pantzanis, 2005), el estudio de la influencia del tamaño de ventana y las discontinuidades de fase (Kirpatrick, 2006) o el análisis de la influencia de la variación de las características espectrales de los formantes (Klabbers, 2007).

2 *Enfoque del problema*

A partir del análisis de los trabajos anteriormente citados, se observa que todavía no se ha conseguido definir una medida que destaque sobre las demás y parece que se empieza a trabajar en otras direcciones de investigación. En este contexto, este trabajo pretende presentar una nueva metodología para seleccionar qué combinación medida-parámetro permite detectar mejor las discontinuidades espectrales. Esta metodología parte de la hipótesis que las distancias con comportamiento más homogéneo (i.e. con media más cercana a 0 y desviación estándar menor) obtenidas al evaluar uniones naturales serán las más eficientes a la hora de detectar discontinuidades.

Esta metodología sigue distintas fases (véase la Figura 1). Primero se realiza un análisis del corpus de voz utilizado basado en: agrupación (*clustering*) fonética y espectral (para calcular la media y la desviación de los parámetros), cálculo de las medidas en estudio empleando la información extraída de la agrupación y el análisis estadístico y la tipificación (también conocida como *z-score*) del comportamiento de las medidas objetivas.

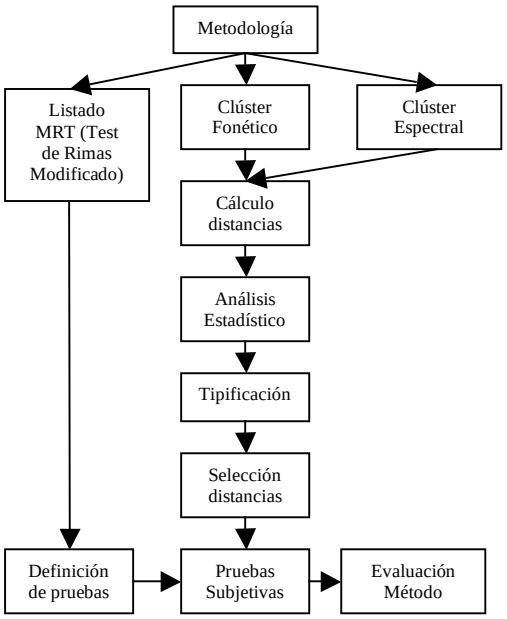

Figura 1: Esquema del proceso seguido en el estudio de las medidas objetivas de estimación de discontinuidades.

Una vez seleccionadas las medidas que presentan un comportamiento más homogéneo, estadísticamente hablando, se procederá a evaluar la hipótesis de partida realizando una serie de pruebas subjetivas sobre un conjunto reducido de monosílabos tipo CVC, donde C indica consonante y V vocal obtenidos de un test de rimas (Stylianou, 2001;Syrdal, 2005). El objetivo es determinar qué distancias objetivas presenta una mayor correlación con los usuarios al estimar la naturalidad de las uniones CVC.

Las distancias consideradas en el estudio son: *i*) Itakura-Saito, con coeficientes FFT y *ii*) Euclídea, Mahalanobis y Donovan, con coeficientes LPC, LSF, información de los tres primeros formantes (frecuencia, ancho de banda y energía) denotada como C3F, MFCC y MFCC con coeficientes delta (MFCC D) y energía (MFCC E) o con ambos (MFCC DE). Este conjunto de parejas distancia-coeficiente cubre la mayoría de los casos presentados en la literatura clásica sobre el tema. Asimismo, el estudio considera las características fonéticas y espectrales del corpus empleado.

3 *Agrupación del corpus*

Dada la dificultad de definir una única distancia como coste de unión para todos los contextos fonéticos en los que se puede encontrar una

unidad acústica, generalmente se opta por organizarlos mediante agrupación fonética y/o espectral (Donovan, 2001; Syrdal, 2005).

En este trabajo se ha utilizado un corpus neutro de voz femenina en catalán, cedido por la UPC, con una duración de 1,5 h. Nótese que la voz femenina permite una tasa de detección de discontinuidades audibles mayor que la voz masculina (Syrdal, 2001). A continuación, se presentan los resultados obtenidos del proceso de agrupación sobre el corpus en estudio.

3.1 Clúster fonético

Según (Syrdal, 2005), el efecto del contexto fonético tiene más influencia a la hora de detectar discontinuidades que la información espectral, por lo que en este trabajo se organiza el análisis de las discontinuidades acústicas según su contexto fonético.

Como primer paso, se agrupan los fonemas del corpus en estructuras CVC según su fonema vocálico, sobre un total de 21654 estímulos. Como se muestra en la Figura 2, el conjunto mayoritario es el que contiene como núcleo vocálico la vocal /@/[1], que está presente en casi la mitad de los estímulos CVC del corpus.

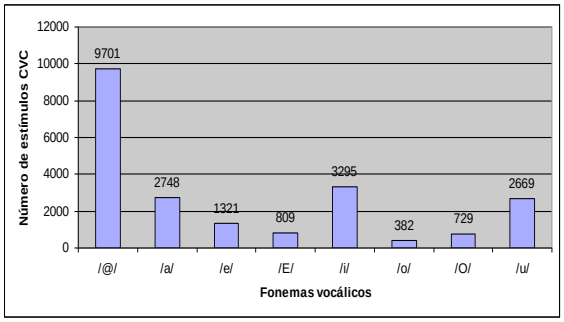

Figura 2: Histograma de la distribución de los estímulos CVC por fonema vocálico.

Al diseñar un corpus de propósito general, generalmente, se tienen en cuenta las características estadísticas de la lengua que trata (i.e. frecuencia de los fonemas), por lo que el corpus suele presentar una buena correlación con la distribución estadística de los fonemas del idioma de trabajo. En este caso, la correlación entre la frecuencia de los fonemas vocálicos en los CVCs extraídos del corpus respeto a la de la lengua catalana de (Rafel, 1979) se obtiene una correlación de ρ=0.99 (véase la Tabla 1).

[1] En este artículo ese emplea notación SAMPA. Véase www.phon.ucl.ac.uk/home/sampa/home.htm

	CVCs	Rafel
/@/	0,45	0,44
/a/	0,13	0,11
/e/	0,06	0,07
/E/	0,04	0,03
/i/	0,15	0,17
/o/	0,02	0,05
/O/	0,03	0,03
/u/	0,12	0,12

Tabla 1: Frecuencia de los fonemas vocálicos en los estímulos CVCs respecto a (Rafel, 1979).

Por otro lado, trabajos previos concluyen que la aparición de discontinuidades espectrales en las vocales depende de su contexto fonético previo y posterior (Syrdal, 2001). Por ello, los estímulos se agrupan considerando el modo de articulación de su contexto consonántico (Syrdal, 2005), así como su sonoridad, ya que la detección de discontinuidades es más elevada en contextos consonánticos sonoros (Syrdal, 2001). Esto es debido a que las consonantes sonoras tienen una fuerte influencia en términos de coarticulación sobre la vocal que las precede.

Por lo tanto, se establecen 8 categorías de CVC según la consonante prevocálica (no se incluye el contexto fonético silencio) y 9 según la postvocálica. Los contextos fonéticos en estudio son: aproximante, fricativa sonora y sorda, lateral, nasal, oclusiva sonora y sorda, vibrante y silencio (sólo para postvocálico). La Figura 3 muestra su distribución en el corpus.

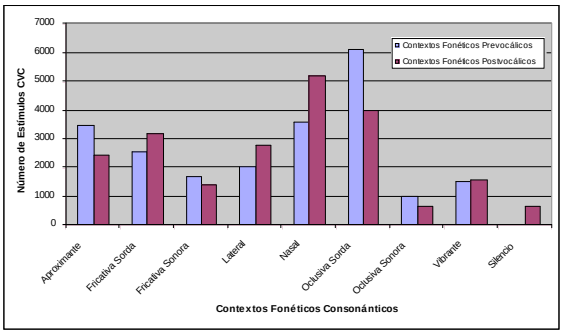

Figura 3: Histograma de la distribución de los estímulos CVC para contextos fonéticos prevocálicos y postvocálicos.

Si se calcula la correlación entre los porcentajes de fonemas consonánticos en los CVCs en estudio respecto a los indicados en (Rafel, 1979), se obtiene una correlación de ρ=0.9 (véase la Tabla 2).

Por lo tanto, de los resultados de correlación obtenidos, se puede concluir que los estímulos considerados son representativos del idioma de

trabajo (i.e. el estudio utiliza información fonéticamente balanceada).

	CVCs	Rafel
Aproximante	0,14	0,10
Fricativa	0,21	0,20
Lateral	0,11	0,12
Nasal	0,20	0,19
Oclusiva	0,27	0,37
Vibrante	0,07	0,11

Tabla 2: Frecuencia de los fonemas consonánticos en los CVCs y en (Rafel, 1979).

4 Análisis de las distribuciones de las medidas objetivas sobre uniones naturales

Cuando se calcula la distancia espectral entre dos difonemas CV-VC procedentes del habla natural, teóricamente su valor debería de ser nulo (o muy cercano a cero). Sin embargo, no todas las combinaciones distancia-parámetro presentan este comportamiento.

Con el objetivo de determinar qué medidas objetivas presentan una distribución de valores con media más cercana a cero y menor desviación típica, se estudia la forma de las distribuciones de las medidas objetivas en estudio sobre *uniones naturales*. Este trabajo parte de la hipótesis que cuanto menos oscile el valor de las distancias respecto a la media en las uniones naturales (idealmente una delta de Dirac), la probabilidad de que la medida objetiva sea un buen detector de discontinuidad aumenta. Del resultado de este análisis se escogerán las combinaciones distancia-parámetro que presenten un comportamiento más cercano al deseado para ser usadas en los experimentos subjetivos.

4.1 Media de las distribuciones

Como primera parte del estudio, se analiza la media de las distribuciones de las medidas objetivas consideradas. Este estudio se ha centrado en los estímulos CVC con vocal /@/, ya que éste es el grupo más numeroso en el corpus, por tanto, de mayor robustez estadística.

En términos de combinación distancia-coeficiente, se observa que la distancia que presenta una media menor es la Euclídea aplicada sobre parámetros LSF (véase la Figura 4). En el otro extremo se encuentra la distancia de Itakura, que es la que presenta la media más alta del conjunto de medidas objetivas estudiado. La distancia de Donovan es la que tiene un comportamiento más estable, independientemente del coeficiente empleado, y suele presentar una media cercana a cero (≅1).

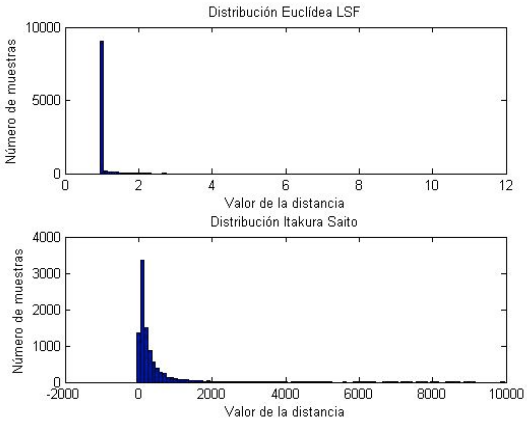

Figura 4: Distribución de la media de las medidas Euclídea-LSF e Itakura Saito sobre los estímulos /C@C/.

4.2 Desviación de las distribuciones

Además de considerar la media de la distribución, se estudia también su desviación (que también debe tender a 0). El problema surge al intentar comparar las distribuciones, ya que éstas presentan distribuciones muy distintas entre sí, según la medida objetiva considerada, para todos los contextos fonéticos analizados.

Por lo tanto, resulta necesario homogeneizar las distribuciones para compararlas correctamente. En este trabajo, se ha optado por aplicar el teorema del límite central (TLC) sobre las distribuciones de partida, para obtener una distribución muestral del valor de la media de la distribución original. Las variables empleadas en el TLC son: 1000 ciclos, que nos garantiza poder calcular con fiabilidad el tercer y cuarto momentos, y 40 muestras/ciclo, para todos los contextos fonéticos (valor único para uniformizar la disparidad de tamaños existente).

Dado que no se consigue el número mínimo de muestras para todos los contextos en todos los fonemas vocálicos en estudio, se decidió agrupar los datos de las vocales /e/+/E/ y /o/+/O/, dada su similitud espectral –al igual que en (Syrdal, 2005), donde no se tiene en cuenta la influencia de la apertura de las vocales en el estudio de las discontinuidades.

La figura 6 presenta la media y la desviación de la simetría o *skewness* (S) y la *kurtosis* (K) de las distribuciones resultantes después de aplicar el TLC. Se puede observar como aparecen dos tipologías distintas de distribuciones. Por un lado, las distribuciones de las vocales /@/ e /i/ tienen forma

leptocúrtica (K>3) y una media estirada hacia la izquierda (S≅1). Por otro lado, se encuentra el resto de vocales, con valores de K y S cercanos a los típicos de las distribuciones gaussianas, cuestión corroborada, mediante la aplicación test de Kolmogorov-Smirnov, con $p<0.05$.

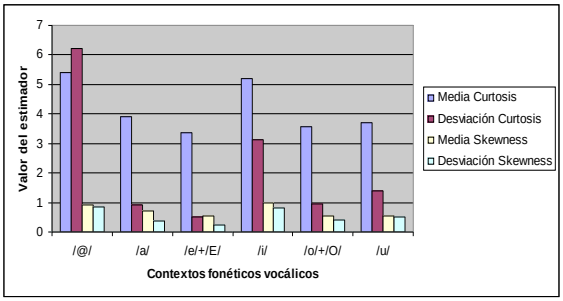

Figura 6: Valor medio y desviación de la simetría y la kurtosis para las distribuciones de las medidas objetivas según fonema vocálico.

5 *Comparativa de las medidas objetivas*

Por un lado, dado que las distribuciones dependen del tipo de coeficiente utilizado, además de aplicar el TLC, resulta necesario tipificar las distribuciones para permitir una comparativa fidedigna de sus desviaciones. Por otro lado, a partir de los resultados observados en términos de momentos de orden tercero y cuarto, resulta necesario definir un único parámetro para evaluar la homogeneidad de las distribuciones alrededor de la media. Cabe comentar que, en una distribución gaussiana, el 68% de los datos se encuentran en el margen definido por su media ± su desviación estándar, obteniendo un valor de tipificación de 1. Tomando este valor como referencia, las distribuciones leptocúrticas (más apuntadas que la gaussiana) tomarán un valor de tipificación <1, por el mayor número de muestras cercanas a la media. Por ejemplo, el fonema /i/ es el que presenta el mayor número de distribuciones leptocúrticas.

Para trabajar con un número razonable de datos (se parte de 22 medidas distancia-parámetros × 17 contextos × 6 vocales), sólo se consideran las 5 mejores combinaciones distancia-parámetro en términos de su valor de tipificación (ordenadas de menor a mayor valor de tipificación) para cada uno de los contextos fonéticos en estudio.

En la figura 7 se muestra el número de contextos para los que cada par distancia-parámetro presenta mejor tipificación en forma de histograma acumulado. Se puede observar como la distancia de Donovan es la que presenta el mejor comportamiento global y que los parámetros LPC, C3F y MFCC (con sus variantes) son los más representados.

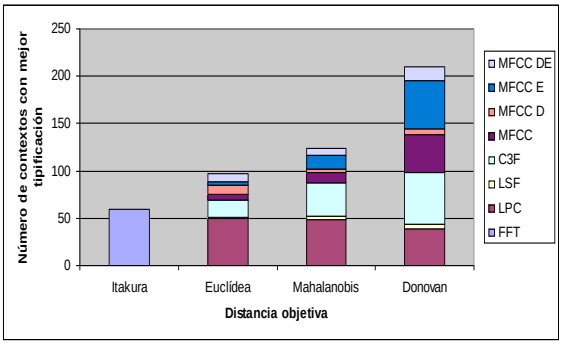

Figura 7: Número de contextos en los que una medida objetiva es de las que mejor tipificación presenta según coeficiente.

Si se analizan las combinaciones distancia-coeficiente en estudio, la que mejor tipificación presenta es la de Itakura-FFT, seguida muy de cerca por la de Donovan-C3F. Las medidas en las que se emplea información (media y varianza de los centroides) procedente del VQ (i.e. distancias Donovan y Mahalanobis) provocan una disminución de la eficiencia de los coeficientes LPC y MFCC D, obteniéndose mejores tipificaciones con la distancia Euclídea para estos coeficientes. Respecto a los coeficientes C3F y MFCC, el comportamiento en términos de tipificación es el inverso al anterior. De la figura 7 se puede concluir que el parámetro LSF no es un buen predictor de la discontinuidad. Finalmente, comentar que la inclusión del coeficiente energía junto a los MFCC tiende a mejorar el valor de tipificación.

6 *Pruebas subjetivas*

Una vez analizadas las medidas objetivas estadísticamente, se procede a estudiar su correlación con la percepción humana. El objetivo de los experimentos subjetivos es contrastar la hipótesis que las distancias con mejor comportamiento en términos de tipificación son capaces de predecir (y modelar) mejor las discontinuidades espectrales.

6.1 Diseño de las pruebas

Siguiendo lo indicado en (Stylianou, 2001; Syrdal, 2005), el diseño del test perceptual parte de un listado de monosílabos tipo CVC procedentes de un Test de Rimas Modificado (MRT), en este caso adaptado al catalán (Alías,

2007). Sin embargo, el diseño de un MRT se realiza de tal forma que abarque el mayor número de monosílabos del idioma, provocando la inclusión de palabras poco habituales en el caso de idiomas con bajo porcentaje de este tipo de palabras, como pasa en catalán (Alías, 2007).

Sin embargo, resultó muy complejo encontrar estímulos CVC del MRT con más de 32 muestras en el corpus, umbral fijado para dotar de fiabilidad estadística a los resultados. Por ello, se decidió escoger para las pruebas los CVC con los contextos fonéticos y los fonemas vocálicos mejor representados (véanse las Figuras 2 y 3). Concretamente, los contextos fonéticos en estudio serán fricativos, nasales y oclusivos. Para la elección de los fonemas vocálicos se tuvieron en cuenta dos criterios: por un lado, el de representación en el corpus, y por otro, el hecho de que al analizar la *kurtosis* y la simetría se observan dos tendencias en las distribuciones de las medidas objetivas sobre uniones naturales: las que presentan /@/ e /i/, con mayor grado de leptocurtismo, y las del resto de fonemas vocálicos con un comportamiento más gaussiano. Por estas razones, los fonemas vocálicos escogidos fueron la /@/ y la /a/. Además, se introdujo el CVC /s@k/ en las pruebas por razones de limitación de corpus (aunque no esté presente en el MRT).

Los estímulos empleados en las pruebas no fueron sintetizados mediante un CTH, sino que se generaron mediante la sustitución del difonema -VC de la estructura CVC por otros difonemas candidatos (emulando el proceso de selección de unidades), manteniendo fijo el difonema CV-. El hecho de no pasar por un proceso de síntesis estrictamente hablando, evita la interferencia del procesado de la señal en el proceso de valoración del comportamiento de las medidas objetivas en estudio.

Los estímulos se presentaron en una frase portadora en la que se sustituía en 32 ocasiones el difonema -VC manteniendo fija la parte CV-:

- Si algú es pen***sa q***ue la comissió - /s@k/ (*Si alguien se piensa que la comisión*)
- Economia i finan***ces*** - /s@s/ (*Economía y finanzas*)
- Mentres***tan***t els nous habitants - /tan/ (*Mientras los nuevos habitantes*)

donde el estímulo CVC a evaluar está marcado en negrita.

6.2 Diseño del experimento

En las pruebas subjetivas participaron 5 evaluadores (3 hombres y 2 mujeres). Como fase previa, se presentó a cada evaluador una serie de estímulos de entrenamiento para que pudiera familiarizarse con el proceso de evaluación, como en (Klabbers, 2001), indicándoles como distinguir la discontinuidad espectral de otros aspectos producidos al insertar el difonema -VC candidato. Las pruebas fueron realizadas usando una interfaz implementada en Matlab utilizando auriculares.

La calificación de la naturalidad de los estímulos CVC sigue la escala MOS de 1 (peor) a 5 (mejor). Los informantes podían escuchar las uniones las veces que necesitaran, pero una vez puntuada la unión no podían volver a evaluarla. Asimismo, tenían la posibilidad de escuchar el estímulo CVC original, la frase portadora original, el CVC generado y la frase portadora que lo incluía. El proceso de pruebas tuvo una duración media de unos 30 minutos.

6.3 Resultados de las pruebas

La evaluación de la capacidad de mapeo subjetivo de las medidas objetivas se obtiene a través de su correlación con la media de las puntuaciones MOS de los informantes. En una situación ideal, la mejor medida debería presentar una correlación ρ=-1, ya que la unión natural (MOS=5) debería darse para una distancia mínima (tendiendo a 0).

Tras la realización de las primeras pruebas, se obtuvieron valores significativos para el estímulo /tan/ (ρ=-0.43 en el caso de Donovan MFCCC y Donovan MFCCE) mientras que los valores de ρ para /s@k/ y /s@s/ fueron inferiores, con máximos de ρ=-0.07 y ρ=-0.14, respectivamente. Nótese que, aunque parezca un valor de correlación bajo matemáticamente hablando, el valor obtenido para /tan/ es cercano al obtenido en experimentos similares recientes (Klabbers, 2007). Aplicando t-student se obtienen valores de confianza del 98% sobre los valores de correlación obtenidos.

Dado que los mejores resultados se obtuvieron para el último estímulo evaluado (/tan/), se decidió estudiar los 2 primeros estímulos de nuevo, partiendo de la hipótesis que a mayor experiencia (con 96 uniones evaluadas) en la realización de las pruebas se consiguen mejores valores de correlación. Tras esta segunda iteración el valor de correlación

mejoró notablemente, con un valor máximo de ρ=-0.35 para Mahalanobis-LSF (con fiabilidad del 95% según t-student) para el estímulo /s@k/. El comportamiento más uniforme se obtiene para los coeficientes MFCC y deltas con valores entorno al -0.3 (véase la Figura 8).

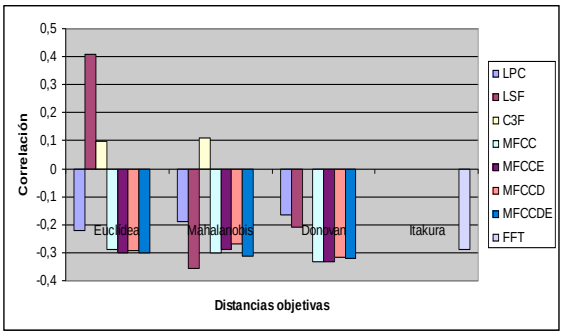

Figura 8: Correlación perceptual para el estímulo /s@k/.

Sin embargo, la correlación obtenida para el estímulo /s@s/ continua presentando valores bajos (véase la Figura 9). Esto puede ser debido a que, según los evaluadores, ésta fue la prueba más difícil de evaluar, al encontrarse el CVC a final de frase. El valor máximo de correlación obtenido es de ρ=-0.31 para la medida Donovan-C3F, pero con patrón de correlación menos estable que en los otros dos estímulos.

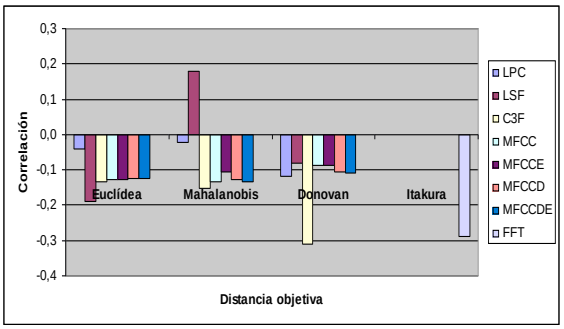

Figura 9: Correlación perceptual para el estímulo /s@s/.

Tras observar que los resultados mejoraban al incrementar la experiencia de los informantes, se decidió hacer también una segunda iteración para /tan/, en la que por cuestiones de disponibilidad sólo participaron 3 informantes. De nuevo, los resultados mejoraron notablemente hasta alcanzar una correlación de ρ=-0.66 para Mahalanobis-MFCCE, con una fuerza estadística del 99.9%. Para el resto de parámetros MFCC y derivados se obtienen valores alrededor de -0.6 (véase la Figura 10).

Una vez comprobado que se obtienen valores de correlación perceptual significativos, se procede a evaluar la viabilidad de la metodología propuesta. Para ello se ordenan las distancias según la tipificación para cada contexto fonético dándole a la medida con mejor tipificación el valor de 22 (igual al número de distancias) y a la peor el valor de 1.

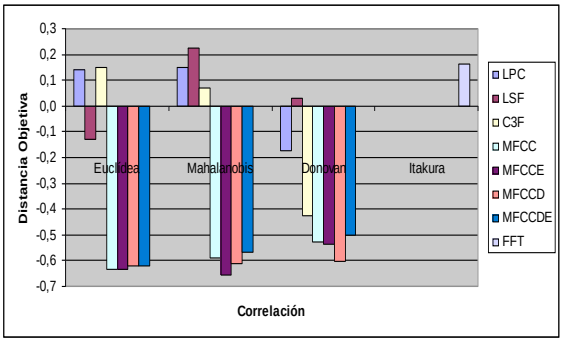

Figura 10: Medidas objetivas con valores de correlación más altos para el estímulo /tan/.

La correlación se calculará entre el valor de correlación obtenido en las pruebas preceptúales y el valor medio resultante de la ordenación por tipificación para los contextos prevocálicos, postvocálicos y el global de contextos para cada fonema vocálico.

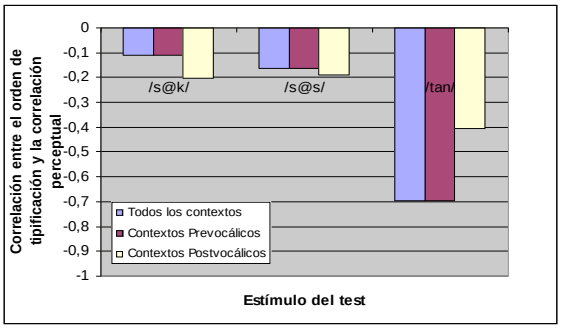

Figura 11: Correlación entre el orden de tipificación y el valor de la correlación perceptual para los estímulos del test.

La Figura 11 muestra como para el estímulo /tan/, se obtiene una ρ=-0,69, mientras que para los otros estímulos se obtiene una ρ<-0,2, siguiendo de algún modo un patrón similar al observado en el estudio anterior. Asimismo, se puede observar como los contextos postvocálicos tienen menor correlación perceptual, resultado inverso al descrito en (Syrdal, 2001). A la hora de calcular las distancias de Mahalanobis y de Donovan para los estímulos del test sólo se consideró la información estadística del centroide respecto al difonema

CV, como esto puede provocar un sesgo hacia los contextos prevocálicos, se decidió recalcular estas distancias considerando la información del centroide del difonema VC. Este estudio se centró en el estímulo /tan/ por ser el más significativo.

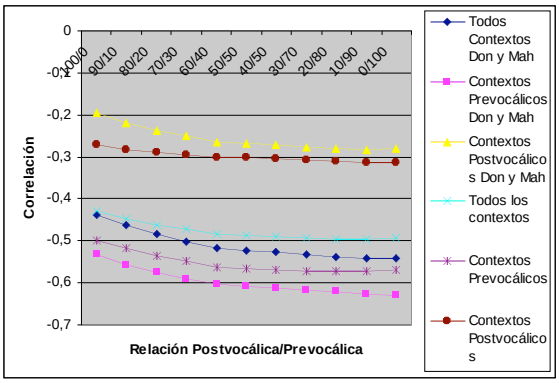

Figura 12: Variación de la correlación con distintos porcentajes de información prevocálica y postvocálica.

La figura 12 confirma que a mayor información prevocálica se obtiene una mayor correlación respecto al orden de tipificación. No obstante, a nivel de correlación perceptual en algunos casos se consiguen valores mayores cuando se incluye mayor información postvocálica (p.ej. Mahalanobis MFCC DE para prevocálica obtiene una ρ=-0.56 y para postvocálica ρ=-0.69).

7 Conclusiones y líneas de futuro

En el presente trabajo se ha propuesto una metodología para estimar la capacidad de predecir la presencia discontinuidades audibles de una serie de medidas objetivas. Esta metodología se basa en la hipótesis que las mejores medidas serán aquellas que presenten una distribución estadística más homogénea, i.e. media cero y mínima desviación estándar, una vez muestreada y tipificada. Esta hipótesis queda validada por los resultados del análisis de las distancias sobre uniones CVC naturales junto a la correlación de las mejores distancias con el test perceptual realizado sobre uniones CV-*, también analizado según el contexto pre y postvocálico. No obstante, resulta necesario seguir trabajando en más pruebas subjetivas para verificar los resultados obtenidos.

Agradecimientos

Los autores quieren agradecer al Dr.Antonio Bonafonte de la Universitat Politècnica de Catalunya la cesión del corpus de voz utilizado.

Bibliografía

Alías F. y M. Triviño 2007. A phoneticallly balanced modified rhyme test for evaluating Catalan speech intelligibility. En *Proc. de ICPhS*, paper 1210.

Donovan R. 2001. A new distance measure for costing spectral discontinuities in concatenative speech synthesis. En *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*.

Hunt A. y A. Black 1996. Unit selection in a concatenative speech synthesis system using large speech database. En *Proc. de ICASSP*, pp. 373–376.

Kirkpatrick, B., D. O'Brien y R. Scaife 2006. Feature extraction for spectral continuity measures in concatenative speech synthesis, En *Proc. de Interspeech*, paper 1385.

Klabbers E., J. van Santen y A. Kain 2007. The contribution of various sources of spectral mismatch to audible discontinuities in a diphone database, En *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), pp. 949-956.

Klabbers E. y R. Veldhuis 2001. Reducing audible spectral discontinuities. En *IEEE Transactions on Speech and Audio Processing*, 9, pp. 39–51.

Pantazis Y., Y. Stylianou, y E. Klabbers 2005. Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis, En *Proc. de Eurospeech*, pp. 2817 - 2820.

Rafel, J. 1979. Dades sobre la freqüència de les unitats fonològiques en català, En *Estudis Universitaris catalans XXIII*, vol 2, 473-496.

Stylianou Y. y A. Syrdal 2001. Perceptual and objective detection of discontinuities in concatenative speech synthesis. En *Proc. de ICASSP*, vol 2, pp. 837-840.

Syrdal A. K. 2001. Phonetic Effects on Listener Detection of Vowel Concatenation, En *Proc. de Eurospeech*, pp. 979-982.

Syrdal A. K. y A Conkie 2005. Perceptually based data-driven join costs: Comparing join types, En *Proc. de Eurospeech*, pp. 2813–2816.

Vepa J. y S. King 2006. Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis. En *IEEE Transactions on Speech and Audio Processing*, 5 (14), pp. 1763- 1771.

Wouters J. y M. Macon 1998. Perceptual evaluation of distance Measures for concatenative speech synthesis. En *Proc. de ICSLP*, pp. 2747–2750.