Evaluación Objetiva y Subjetiva de Entonación Sintética

David Escudero Mancebo, César González Ferreras,
Valentín Cardeñoso Payo

Departamento de Informática
Universidad de Valladolid
{descuder,cesargf,valen}@infor.uva.es

Resumen

En esta comunicación se presentan una serie de tests de evaluación de entonación sintética. Los tests han sido aplicados sobre los resultados obtenidos empleando una metodología de modelado de entonación basada en funciones de Bézier y modelado estadístico con aplicaciones en conversión texto voz. Primero se describen una serie de generalidades sobre tests objetivos y subjetivos, apuntando las ventajas e inconvenientes de unos frente a otros. Después se define un test objetivo y se hace una valoración de los resultados obtenidos. El test objetivo se completa con un test percetual poniendo de manifiesto la calidad de la entonación sintética generada [1] .

1. Introducción

Los esfuerzos en mejorar la entonación de los sistemas traductores texto-voz, deben llevar asociados un conjunto de métodos y técnicas que permitan evaluar las mejoras realizadas. Existen métodos de evaluación objetivos y subjetivos para valorar las mejoras en la entonación, que aportan beneficios tanto a diseñadores como a posibles usuarios.

Desde el punto de vista de los diseñadores, la evaluación es una herramienta para cuantificar las mejoras realizadas y hacer diagnósticos de los posibles aspectos por mejorar. Desde el punto de vista del usuario o posible cliente, las técnicas de evaluación le permitirán comparar sistemas para poder decidirse a la hora de hacer una posible elección.

Se han descrito tanto métodos subjetivos como objetivos de evaluación de entonación. Cada uno de ellos aporta un valor que hace que, en el estado del arte actual, no podamos prescindir de ninguno de ellos. A continuación, se describen los distintos métodos de evaluación, comentando sus ventajas e inconvenientes.

1.1. Evaluación Subjetiva de Entonación Sintética

La idea básica de estos métodos se describe en los trabajos clásicos del IPO [1] bajo el nombre de *analysis by synthesis*. El método consiste en reproducir una locución empleando una curva de entonación sintética y preguntar a un usuario sobre la calidad o naturalidad de la misma. En [2] o [3] puede encontrarse un compendio de métodos y técnicas subjetivas para evaluar distintos aspectos de voz sintética, entre otros, de la entonación. Básicamente existen dos técnicas: la comparación de pares y la puntuación de locuciones.

En la comparación de pares, se emiten dos locuciones en condiciones distintas (p.e. una con $F0$ original y otra con $F0$ sintético) y un usuario compara las mismas. El usuario debe decir cual prefiere y justificar su decisión. El principal problema de este método es la dificultad de encontrar un usuario capaz de emitir juicios válidos sobre los parecidos de las locuciones. Otro problema es que la prueba se ha de limitar a locuciones breves, porque es difícil que el usuario recuerde los detalles en locuciones mayores. Este método tiene especial interés cuando es realidado por un usuario bien entrenado, ya que éste puede emitir diagnósticos que localicen aspectos concretos susceptibles de mejora.

En el método de puntuación, el usuario escucha una locución y le asigna una nota numérica en función de cómo de natural le parezca. Estos tests son más eficaces para locuciones largas. El usuario no necesita recordar cómo fueron las locuciones anteriores, sino que sólo debe poner una nota final. El problema es que si se aplica a frases cortas, el usuario puede no tener suficientes elementos de juicio.

Un problema de estos métodos, es que los juicios de los usuarios pueden estar condicionados por otros aspectos distintos a los que se quiere evaluar. Para evitar esto, es importante que otros aspectos que no son entonación, como puede ser la calidad de la voz, o la duración de segmentos, sean lo más parecido posible a los que se observan en las locuciones reales.

Además, los métodos de evaluación subjetivos son fuertemente dependientes del usuario empleado para emitir los juicios. Dos usuarios diferentes pueden emitir juicios muy distintos sobre las mismas locuciones. Para paliar estos efectos, se eligen usuarios pertenecientes a ámbitos sociológicos similares. Los oyentes deberán pertenecer, a ser posible, al tipo de usuario que potencialmente llegaría a emplear el sistema.

La subjetividad de los oyentes, aunque de hecho es un inconveniente en este tipo de test, también puede ser vista como un valor importante. Hay que tener en cuenta que los usuarios de conversores texto-voz, cuando deben tomar la decisión de utilizar o comprar un sistema texto-voz emplean, por lo general, métodos subjetivos.

Los métodos subjetivos son costosos en el tiempo porque implican la movilización de un grupo de evaluadores. Además, no son repetibles, en el sentido de que con el mismo grupo de usuarios y realizando los mismos tests pueden obtenerse resultados diferentes. Para el día a día en el desarrollo de modelos de entonación, se hace necesario emplear tests objetivos realizables automáticamente y que sean un indicador de las posibles mejoras que se vayan introduciendo.

1.2. Evaluación Objetiva de Entonación Sintética

La idea básica de los métodos objetivos de evaluación de entonación es medir la diferencia entre la entonación natural y la entonación sintética. El objetivo es calcular la distancia entre

[1] Este trabajo ha sido parcialmente financiado por el proyecto VA16-00 de la Consejería de Educación de Castilla y León y por el proyecto C2000-1669-C0403 de la CICyT

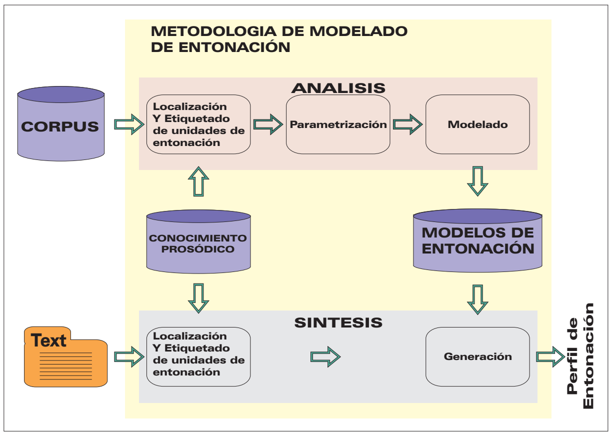


Figura 1: Metodología de modelado y generación de entonación.

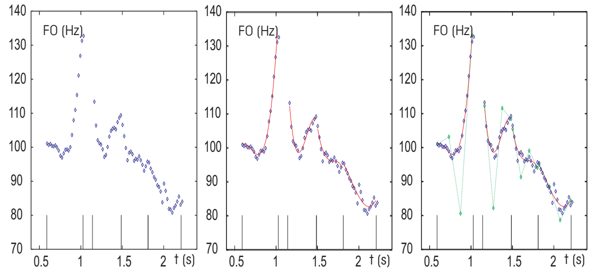


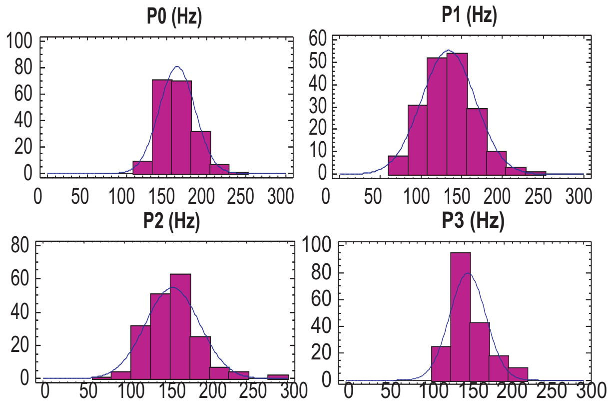
Figura 2: Proceso de parametrización: A la izquierda el perfil de entonación de entonación original con marcas de frontera entre grupos acentuales; En el centro la función de Bézier aproximante; A la derecha los puntos de control de la función de Bézier. Los puntos de control son los parámetros de entonación.

un perfil de entonación natural y el perfil sintético correspondiente. Para ello, se han descrito un número de métricas que se enumeran a continuación:

- El RMSE (*root-mean-square error*) mide la distancia entre dos contornos en el eje de tiempos. Mide la distancia entre aspectos particulares de los perfiles de entonación. Cuanto mayor sea el RMSE mayor es la diferencia entre los dos perfiles.
- El *coeficiente de correlación de Pearson* R^2 entre el perfil de pitch original y el sintético, mide el grado de similitud de la evolución temporal de estos perfiles. Este coeficiente establece si dos variables están relacionadas linealmente. Si hay una relación en tiempo y frecuencia entre los dos perfiles, eso significa que su evolución es similar.

Estas dos métricas han sido empleadas en diversos estudios sobre entonación (p.e. [4] [5]). Dik Hermes en [6] hace una revisión de las métricas propuestas y afirma que estas dos métricas son las mejores que se han planteado. A partir del estudio de Hermes aparecen otras propuestas de métricas que pretenden conjugar criterios perceptuales en las métricas objetivas:

- La *mean-absolute-frecuency-deviation* es propuesta por Bellegarda en [7] para determinar como de bien los modelos de pitch capturan las cualidades de los datos originales. Según el autor, las métricas RMSE y R^2 se basan en la norma L_2 que tiende a reducir los efectos de realizaciones raras (*rare outliers*). Una realización rara puede provocar efectos perceptuales significativos y puede ser beneficioso intentar evitarlas. Empleando el valor absoluto se emplea la norma L_1 y se evita el riesgo de minorizar la importancia de las realizaciones raras.
- Campione et al en [8] emplean el error cuadrático medio entre *puntos objetivo* del modelo INTSINT en semitonos. Se apunta como criterio de calidad la proporción de puntos objetivo que se desvían menos de 2 semitonos entre el contorno original y sintético.
- Clark et al. en [9] introducen otras métricas que intentan considerar no sólo la distancia en los valores de frecuencia sino también los desplazamientos temporales de los perfiles de F0. En la referencia se concluye que no se obtienen resultados mejores que con el RMSE y R^2.



	μ	σ
P0	166	17
P1	137	26
P2	162	27
P3	157	19

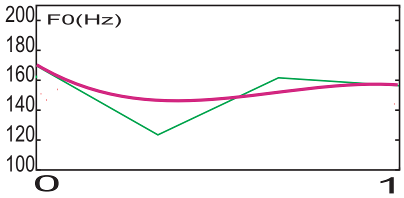


Figura 3: Modelo estadístico inferido para una clase dada. Arriba los histogramas que muestran la distribución de valores de cada parámetro. Abajo a la izquierda los estadísticos de primer orden para la clase. Abajo a la derecha la representación gráfica del patrón característico de la clase.

La aplicación de estas métricas sobre los resultados que se obtienen al aplicar la metodología de modelado que se describe en el apartado siguiente servirá para obtener una evaluación objetiva de los mismos. A la vez, el uso de estas métricas servirá para poner en comparación los resultados obtenidos con respecto a los resultados que obtienen otros autores en trabajos similares.

2. Metodología de Modelado de Entonación

La metodología de modelado de entonación empleada ha sido descrita en [10, 11, 12] y se resume en la figura 1. Los modelos de entonación se generan a partir de un corpus. Primero se localizan en el corpus los grupos acentuales y se enriquecen

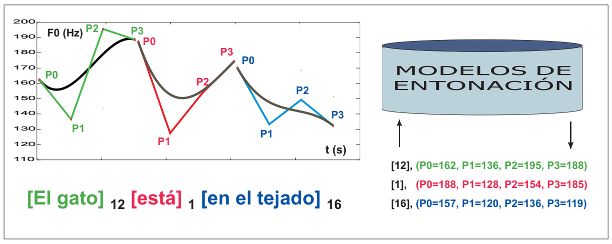


Figura 4: Esquema de generación de perfiles de entonación. Se genera un perfil de F0 para la frase de ejemplo *El gato está en el tejado*.

con atributos prosódicos. Después los perfiles de F0 asociados a cada grupo acentual son parametrizados. Los parámetros de los grupos acentuales sirven para modelar la entonación del corpus. En síntesis, se localizan primero los grupos acentuales en el corpus y se etiquetan prosódicamente. Con estas etiquetas prosódicas se toma el modelo de entonación correspondiente y se generan perfiles sintéticos por simulación de los modelos obtenidos a partir del corpus de estudio.

La parametrización de entonación se hace empleando funciones de Bézier. El perfil de entonación se segmenta en unidades de entonación (grupos acentuales) y cada grupo acentual es parametrizado individualmente. Los puntos de control de la función de Bézier que aproxima el perfil de entonación son los parámetros de entonación del grupo acentual. Los contornos de pitch son aproximados empleando mínimos cuadrados. Se emplean cuatro puntos de control por grupo acentual. En [10] se describe el método de aproximación y las distintas opciones de parametrización que fueron consideradas.

Para modelar se establecen primero clases de grupos acentuales. Las clases se establecen en función de una serie de factores prosódicos. Dos grupos acentuales están en la misma si tienen los mismos valores para los factores prosódicos. Los factores prosódicos se obtienen del trabajo de Garrido [13]. El modelo de entonación para cada clase es la distribución estadística de los parámetros observada en el corpus. El objetivo de seguir este procedimiento es el de modelar la regularidad de los patrones de entonación en la clase, y la variabilidad observada. Para más información sobre el proceso de modelado ver [11]. La gráfica 3 ilustra el contenido de los modelos de entonación.

Los modelos estadísticos obtenidos son empleados para generar nuevos perfiles sintéticos. Para ello, primero se localizan y clasifican los grupos acentuales del texto a sistetizar. Con la clase, se accede a la base de modelos para generar los parámetros correspondientes. Estos parámetros son empleados después para generar el perfil de pitch correspondiente, que será la curva de Bézier generada con los puntos de control como se describe en la figura 4. La forma de obtener los parámetros empleando los modelos estadísticos es por simulación. Para ver una descripción detallada de los métodos de simulación empleados, ver [14].

3. Evaluación de Resultados

En este apartado se describen los resultados de los tests de evaluación realizados para validar la metodología de modelado y generación de entonación sintética. Primero se comentan los tests de evaluación objetiva, describiendo las métricas empleadas y comentando los resultados obtenidos poniéndolos en comparación con los obtenidos en otros estudios. Después se comentan las pruebas perceptuales que se han realizado describiendo los resultados de las mismas.

Corpus	#	Mean	Sigma	RMSE	MAFD	Corr	PS2
CPT	675	157.77	23.41	10.40	7.24	0.89	0.08
CPP	169	157.67	23.73	10.42	7.25	0.89	0.07

Tabla 1: Comparativa de los corpus de prueba. Estadísticos de primer orden y distancias entre los perfiles parametrizados y los originales.

Simulacion	RMSE	MAFD	Corr	PS2
SPM	17.85	13.91	0.73	0.28
SPV	18.53	14.43	0.70	0.30
SVM	25.10	20.31	0.00	0.49

Tabla 2: Resultados en predicción con CPT.

3.1. Evaluación Objetiva

En el apartado 1.2 se hace una revisión de las métricas empleadas para valorar de forma objetiva la calidad de la entonación sintética. Aquí se aplican dichas métricas para evaluar la calidad de los perfiles de F0 sintéticos generados con la metodología expuesta en este trabajo de investigación.

Se emplean como métricas de calidad el **RMSE** (*root-mean-square-error*), el **MAFD** (*mean-absolute-frecuency-deviation*), el coeficiente de correlación **Corr** y la proporción de puntos de F0 que se desvían más de 2 semitonos en valor absoluto **PS2**. Estas métricas se aplican para medir la distancia entre un perfil de F0 de una frase dada y el correspondiente perfil de pitch simulado.

Se emplean dos corpus de prueba. El primero de ellos está formado por todas las frases del corpus y nos referiremos a el como **CPT**. El segundo de ellos se compone del 25 % de las frases del corpus elegidas aleatoriamente y nos referiremos a él con el nombre de **CPP**. Para realizar los test con CPT, se obtienen los modelos de entonación con los grupos acentuales de todo el corpus. Para hacer los tests con CPP, los modelos de entonación se generan con el 75 % de las frases del corpus que no están en CPP.

La tabla 1 pone en comparación los dos corpus de prueba. # es el número de frases del corpus, $Mean$ y $Sigma$ son los estadísticos de los F0 en los corpus. RMSE, MAFD, Corr y PS2 son las métricas ya mencionadas. Los valores de estas métricas

Simulacion	RMSE	MAFD	Corr	PS2
SPM	18.93	14.97	0.70	0.33
SPV	20.43	15.97	0.67	0.35
SVM	25.62	20.80	0.00	0.50

Tabla 3: Resultados en predicción con CPP.

Simulacion	RMSE	MAFD	Corr	PS2
SPM	15.90	12.69	0.77	0.25
SPV	17.60	13.78	0.74	0.28
SVM	23.13	18.92	-0.00	0.47

Tabla 4: Resultados en predicción con CPP con perfiles suavizados.

se calculan al comparar los puntos de los contornos de F0 de las frases sin aplicar ningún suavizado, y los que se obtienen al parametrizar los contornos. Al aplicar estas métricas sobre ambos corpus, los resultados son similares.

Se evalúan tres métodos de generación de perfiles sintéticos: Simulación de Patrones Medios **SPM**, que genera los perfiles de F0 con los valores medios de los parámetros en cada clase. Simulación de Patrones Variables **SPV**, que aplica un método de generación de datos basado en el método de *Thomson-Taylor* (ver [15]) con rechazo de las muestras que se desvía por encima de un umbral del patrón medio; y Simulación de Valor Medio **SVM** donde se generan perfiles de F0 planos (la frecuencia fundamental se iguala al valor medio de F0 observado en el corpus). Este último método se aplica para tener una referencia de perfiles sintéticos de naturalidad nula.

Las tablas 2 y 3 muestran los resultados de aplicar las métricas sobre los corpus de prueba CPT y CPP respectivamente. Al aplicar simulación y comparar los resultados se observa que las métricas con SPM son mejores que las de SPV y que ambos son mucho mejores que SVM.

Los resultados en la tabla 2 son ligeramente mejores que los resultados en la tabla 3. Este hecho era previsible ya que en CPT, al contrario de lo que ocurre con CPP, los datos de prueba están incluidos en los datos de entrada de los modelos.

Los resultados mejores se obtienen para SPM. Esto era previsible porque el valor medio siempre minimiza los errores. Tanto SPM como SPV son mejores que SVM, lo que pone de manifiesto que las mejoras en la naturalidad que se consiguen son evidentes. SPV está más cerca de SPM que de SVM, lo que es un indicador de la calidad de los perfiles obtenidos por simulación.

Estudio	RMSE	Corr
Dusterhoff & Black [16]	32.5	0.6
Ross & Ostendorf [17]	33	No dado
Black & Hunt [18]	34.8	0.62
Dussterhoff F2B [5]	34.3	0.60
Dussterhoff KDT [5]	9.1	0.74
Dussterhoff FHL [5]	21.1	0.53
Moehler & Conkie [19]	32.1	No dado

Tabla 5: Comparación con otros trabajos de investigación sobre generación de entonación.

En la tabla 4 se muestran los valores de las métricas al aplicar el corpus de prueba CPP, pero esta vez haciendo un suavizado de los perfiles de F0 antes de compararlos. Los resultados en la tabla 3 son similares a los resultados de la tabla 4. Al suavizar los contornos de pitch, los resultados son mejores. La explicación de este hecho hay que buscarla en que al filtrar disminuye la varianza de los contornos y con ella el número de puntos espurios y esto hace que las métricas sean mejores. Es de destacar la mejora de la métrica **P2S**. Esto sucede porque al suavizar se elimina una proporción importante de puntos espurios, que pueden superar el umbral establecido de 2 Semitonos (St). Dado que el efecto de los puntos espurios aislados no suele ser perceptible, puede concluirse que para valorar el efecto de esta métrica, es necesario aplicar algún tipo de suavizado sobre los perfiles de F0.

La tabla 5 muestra los resultados obtenidos en otros estudios similares de modelado de entonación. Aunque los datos no son comparables por tratarse de estudios que emplean corpus diferentes, sí sirven de referencia para valorar este trabajo frente a otros del estado del arte.

Por último, hacer una reflexión sobre el valor de PS2. Para el mejor de los casos, el 25 % de los puntos de los perfiles generados se desvía 2 St de los perfiles observados. Aunque este resultado podría mejorar si sólo se comparan las zona vocálicas, más importantes desde el punto de vista de la percepción de la entonación, hay que asumir que los perfiles generados pueden ser perceptualmente diferentes de los que hay en el corpus. Este resultado puede estar justificado por la variabilidad en las clases de entonación. Esta variabilidad se asumió a la hora de definir la metodología de modelado, se observó en los modelos obtenidos, se intentó reproducir, y ahora se refleja en las curvas sintéticas generadas. No hay que interpretar pues este resultado como un mal funcionamiento de la metodología de modelado. No obstante, a la vista de este resultado, se plantea la cuestión de la naturalidad de estos perfiles sintéticos.

Los resultados de los tests objetivos muestran que se generan perfiles similares a los encontrados en los corpus de prueba. También muestran que se están generando perfiles de entonación variables que pueden ser perceptualmente diferentes a los del corpus. Para saber hasta qué punto estos perfiles sintéticos se perciben como contornos de pitch naturales, es necesario utilizar tests perceptuales.

3.2. Evaluación Subjetiva

El objetivo de los tests perceptuales es evaluar la naturalidad de la entonación sintética. Los perfiles de entonación generados con el procedimiento descrito en el apartado 2, pueden ser diferentes para una misma clase de grupo acentual. Esto implica que una misma frase podría tener asociados diferentes perfiles de pitch. Los tests perceptuales servirán para evaluar el grado de naturalidad con el que se perciben diversos perfiles asociados a una misma frase.

Para hacer un test perceptual se puede optar por dos opciones: emplear un sistema conversor texto-voz para generar las locuciones a puntuar o emplear resíntesis PSOLA [20]. Aquí sólo se han realizado los tests perceptuales con resíntesis PSOLA. Aunque los modelos se han incluido en los sistemas conversores texto-voz Parlante [21] y MLTTSUPC [22], no se ha empleado esta opción para realizar los tests. La razón es que emplear resíntesis PSOLA permite generar locuciones sintéticas en las que lo único que varía con respecto a una locución original es el pitch. Empleando PSOLA, se puede garantizar que la calidad del resto de parámetros de voz: calidad, duración de segmentos etc... permanece inalterado con respecto a la locución natural. De este modo, el oyente que evalúa las locuciones sintéticas no va a condicionar su decisión a otros factores que no son los perfiles de pitch. Se utiliza el software `praat` [2] para realizar la resíntesis PSOLA.

[2] **praat** es un software freeware para trabajos de fonética experimental `www.praat.org`

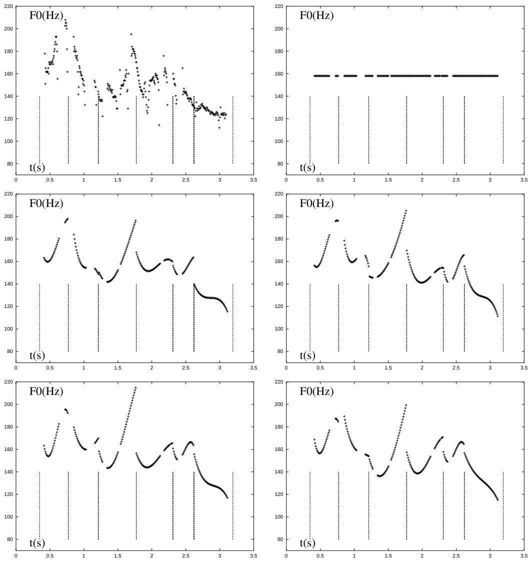


Figura 5: Diversos contornos de $F0$ generados para la misma frase en el test perceptual. Arriba a la derecha está el perfil original y abajo a la izquierda un perfil de F0 plano. Estos perfiles sirven para establecer la referencia de entonación perfecta y de entonación nula respectivamente. La frase empleada es ”Con ese caracter abnegado ahuyentas a todo el mundo” (referencia esmas0146 del corpus ESMA-UPC). Las marcas verticales indican las fronteras de grupos acentuales: [Con ese][caracter] [abnegado][ahuyentas][a todo][el mundo].

3.3. Tests perceptuales

El objetivo del test es evaluar la naturalidad de un conjunto de contornos de entonación variables generados para una misma frase. El método de evaluación empleado se basa en el descrito por Ostendorf et al. en [23]. Doce oyentes puntuan una serie de locuciones sintéticas. Se eligen tres frases del corpus, y para cada frase, el oyente escucha seis locuciones: la versión original (puntuación=10, naturalidad máxima), una versión sintética en la que el pitch es plano (puntuación=0, naturalidad mínima), y cuatro locuciones sintéticas donde los contornos de pitch generados se crean por simulación. Los oyentes que realizan el test deben puntuar las locuciones con una nota del 0 al 10 de acuerdo a su propio criterio subjetivo. La figura 5 ilustra la forma de los perfiles de F0 sintéticos empleados en este test. La figura 6 muestra las puntuaciones otorgadas por cada oyente a cada una de las locuciones.

En la figura 6 se observa que el rango de variación de las puntuaciones de los oyentes es elevado. Como cabía esperar por tratarse de un test perceptual, las puntuaciones dependen fuertemente de los oyentes. En la tabla no se observa que los oyentes puntúen mejor o peor, de forma regular, determinadas locuciones. Esto pone de manifiesto que perfiles variados pueden percibirse con un grado similar de naturalidad. Por otro lado, hay que poner de manifiesto los valores altos de las puntuaciones obtenidas.

3.4. Síntesis con conversores texto-voz

A pesar de que el test perceptual se ha realizado con resíntesis PSOLA, también se ha incluido el modelo en los conversores

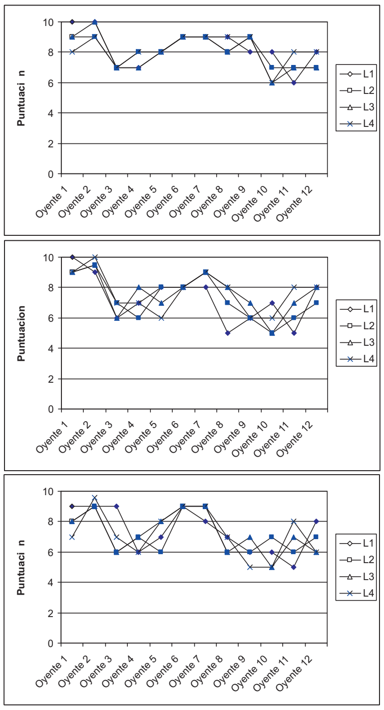


Figura 6: Resultados de la evaluación perceptual. Arriba la Frase 1, en el centro la Frase 2 y abajo la Frase 3 del test

texto-voz Parlante 2.0 y MLTTSUPC [14] y se han comprobado los resultados que se obtienen.

Para escuchar los perfiles sintéticos en el sistema conversor texto-voz **UPCMLTTS**, se incluyen los modelos de entonación en el procedimiento descrito en [14]. No se ha realizado ningún test de tipo perceptual sobre los contornos generados, aunque la sensación es que la entonación del habla generado mejora con respecto a la que se obtiene aplicando el modelo de entonación de que disponía originalmente este conversor.

También se ha incluido el modelo en la arquitectura de conversión texto-voz **Parlante** aplicando el método descrito en la sección [14]. A diferencia de lo que ocurre con el sistema UPCMLTTS, ahora es necesario primero escalar los perfiles de F0. En UPCMLTTS la voz empleada para modelar la entonación y la voz empleada para generar voz es la misma, por lo que no es necesario alterar los perfiles de pitch sintéticos para que puedan ser aplicados en el sistema. Por contra, en Parlante la voz del conversor texto-voz no coincide con la voz que se emplea para elaborar los modelos de entonación. Éste hecho, y el hecho de disponer de un modelo de duración mucho más pobre que el del sistema *UPCMLTTS*, hace que la calidad de las locuciones sea peor.

4. Conclusiones

En esta comunicación se ha puesto de manifiesto la necesidad de disponer de un juego de métricas objetivas que permitan valorar las aportaciones del modelado de entonación propuesto. En el estado del arte actual, las métricas $RMSE$ y R^2 parecen ser las más adecuadas porque ofrecen resultados realistas y porque su uso es el más extendido. Otras métricas que consideran criterios perceptuales son también consideradas. La métrica ideal, que permita obtener de forma automática una comparativa fiable de las diferencias perceptuales provocadas por cambios en los perfiles de pitch aún no ha sido definida. Mientras no exista tal métrica, los tests perceptuales son imprescindibles. Sin embargo, la no repetibilidad y el coste de este tipo de tests parece aconsejar la realización de tests informales frente a pruebas exhaustivas mucho más rigurosas.

Se ha mostrado que el valor de las métricas es sensible al corpus empleado y al suavizado previo de los perfiles de F0 en las muestras del corpus. Cuando los datos del corpus de prueba se utilizan también para modelar la entonación, los resultados mejoran frente al caso en el que los corpus de modelado y de prueba están separados. Al suavizar los contornos de pitch en los datos de prueba, los resultados mejoran, ya que se eliminan las fluctuaciones espurias que se observan en los perfiles de F0.

Emplear los modelos de entonación para generar los contornos de pitch, frente a emplear un modelo simplista de curva de entonación plana mejora sensiblemente todas las métricas de calidad. Este es un indicador claro que pone de manifiesto las mejoras introducidas al aplicar la metodología de trabajo.

Las métricas RMSE y Corr podrían ser empleadas para comparar los resultados aquí obtenidos con los obtenidos en trabajos similares realizados por otros autores. Sin embargo, el hecho de que cada autor trabaje con su propio corpus, limita las conclusiones que estas comparativas podrían aportar.

Los tests perceptuales realizados y la inclusión de los modelos de entonación en sistemas de conversión texto-voz reales, ponen de manifiesto la viabilidad del uso de resultados de este trabajo de investigación con aportaciones en la naturalidad de la entonación sintética. Perfiles generados aplicando un método de simulación se perciben con diferencias apreciables, pero no es posible afirmar que unas realizaciones se perciban con peor calidad que otras. Este hecho pone de manifiesto la posibilidad de generar perfiles sintéticos variables para atenuar la sensación de monotonía que se produce al generar perfiles siempre los mismos perfiles ante idénticos factores prosódicos.

5. Referencias

[1] J. Hart, R. Collier, and A. Cohen, *A perceptual study of intonation. An experimental approach to speech melody*, Cambrige University Press, 1990.

[2] R. Bezooijen V. van Heuven, "Quality Evaluation of Synthesized Speech," in *Speech Coding and Synthesis*, chapter 21, pp. 707–738. Amsterdam: Elsevier, 1995.

[3] A. Acero, X. Huang, and H. HsiadWuen, *Spoken Languaje Processing*, Carnegie Mellon University, 2001.

[4] K. Ross, *Modeling of intonation for Speech Synthesis*, Ph.D. thesis, College of Engineering, Boston University, USA, 1994.

[5] K. E. Dusterhoff, *Synthesizing Fundamental Frequency Using Models Automatically Trained from Data*, Ph.D. thesis, University of Edimburgh, U.K., 2000.

[6] D. J. Hermes, "Measuring the perceptual similarity of pitch contours," *Journal of Speech, Language, and Hearing Research*, vol. 41, pp. 73–82, February 1994.

[7] J. R. Bellegarda, K. Silverman, and V. Anderson, "Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation," *IEEE Transaction on Speech and Audio Processing*, vol. 9, no. 1, pp. 52–66, January 2001.

[8] E. Campione and Véronis J., "A statistical study of pitch target points in five languages," in *Proceedings of ICSLP 98*, 1998.

[9] R. A. J. Clark and K. E. Dusterhoff, "Objetive methods for evaluating synthetic intonation," in *Proceedings of Eurospeech 99*, September 1999.

[10] D. Escudero and V. Cardeñoso, "Corpus based extraction of quantitative prosodic parameters of stress groups in spanish," in *Proceedings of ICASSP 2002*, Mayo 2002.

[11] D. Escudero, C. González, and V. Cardeñoso, "Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis in spanish," in *Proceedings of ICSLP 2002*, Mayo 2002.

[12] V. Cardeñoso and D. Escudero, "Statistical modelling of stress groups in spanish," in *Proceedings of Prosody 2002*, 2002.

[13] J. M. Garrido, *Modelling Spanish Intonation for Text-to-Speech Applications*, Ph.D. thesis, Facultat de Lletres, Universitat de Barcelona, España, 1996.

[14] D. Escudero, *Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión Texto Voz.*, Ph.D. thesis, Dpto. de Informática, Universidad de Valladolid, España, 2002.

[15] J. E. Gentle, *Random Numbers Generation and Monte Carlo Methods (Statistics and Computing)*, Springer, 1998.

[16] K. Dusterhoff and A. Black, "Generating f0 contours for speech synthesis using the tilt intonation theory," in *Proceedings of ESCA Workshop of Intonation*, September 1997.

[17] K. Ross and M. Ostendorf, "A dynamical system model for generating f0 for synthesis," in *Proceeding ESCA Workshop On Speech Synthesis*, 1994.

[18] A. W. Black and A. J. Hunt, "Generating f0 contours from tobi labels using linear regression," in *Proceedings of ICSLP 96*, 1996.

[19] G. Mohler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proceedings of 3 ESCA Workshop on Speech Synthesis*, 1998.

[20] E. Moulines and W. Verhelst, "Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech," in *Speech Coding and Synthesis*, chapter 15, pp. 519–555. Amsterdam: Elsevier, 1995.

[21] L. Feal, D. Escudero, and V. Cardeñoso, "Un modelo arquitectónico para los sistemas texto-voz," in *Actas de las IV Jornadas de Infomatica. Las Palmas de Gran Canaria*, Junio 1998.

[22] A. Bonafonte, A. Febrer, A. Moreno, J.A. Rodríguez, and A. Sesma, "Actividades en el área de conversión de texto a habla en el centro talp," in *Actas de las I Jornadas en Tecnologias del Habla. Sevilla 2000*, Noviembre 2000.

[23] I. Bulyko, M. Ostendorf, and P. Price, "On the relative importance of different prosodic factors for improving speech synthesis," in *Proceedings of ICPhs 99*, 1999.