# MÉTODO SUBJETIVO PARA EVALUAR LA ENTONACIÓN SINTÉTICA

*Eva Navas Cordón* *Juan María Sánchez* *Eluska Sukia Arruabarrena* *Inmaculada Hernáez Rioja*

Departamento de Electrónica y Telecomunicaciones
Universidad del País Vasco/ Euskal Herriko Unibertsitatea
eva, ion, eluska, inma@bips.bi.ehu.es

**ABSTRACT**

This paper describes a method for assesing the quality of synthetic intonation using a subjective method. The perceptual method of evaluating intonation, not only evaluates the quality of synthetic intonation, but also allows us to compare different models of intonation.

## 1. INTRODUCCIÓN

Uno de los problemas que más preocupa actualmente en el desarrollo de los sistemas de conversión de texto (CTV) a voz es la evaluación de la calidad de la señal sintética obtenida. El interés de obtener sistemas fiables de evaluación es múltiple y la metodología de evaluación utilizada depende en gran medida de los objetivos de dicha evaluación. Por un lado, la obtención de una medida de calidad global del sistema de CTV proporciona un elemento de valoración al usuario final ('benchmarking'). Este tipo de evaluación se conoce también como sistema 'de caja negra' y proporciona valoraciones generales o globales. Por otro lado, tiene interés como herramienta de diagnóstico, útil sobre todo para diseñadores y desarrolladores, permitiendo la detección y localización de fallos, y la evaluación por separado de los diferentes módulos que componen el sistema CTV.

El método que nos ocupa en este trabajo pertenece a esta última categoría, y está enfocado a la obtención de una medida de la evaluación de la calidad de la entonación lograda con un determinado modelo entonativo, tratando de aislar en la medida de lo posible este factor de otros aspectos de la señal, tales como su inteligibilidad segmental. Además este método permite comparar dos modelos entonativos diferentes, para saber cuál resulta más natural desde el punto de vista perceptual.

El siguiente apartado describe los objetivos que se pretendían con el desarrollo de este trabajo. A continuación en el punto 3 se describe el proceso de evaluación seguido, para presentar finalmente los resultados en el punto 4.

## 2. OBJETIVOS DE LA EVALUACIÓN

Como ya se ha indicado, la evaluación de la calidad un sistema de CTV puede realizarse de muy diferentes formas en función de cuál sea el objetivo de la misma. En nuestro caso, la evaluación a desarrollar debía proporcionarnos respuesta a las siguientes preguntas:

- ¿Qué grado de similitud mantiene el modelo de entonación elegido con la entonación natural?
- Considerando que se disponía de otro modelo entonativo muchísimo más simple (modelo de Picos y Valles, [1]), y que el desarrollo del nuevo modelo ha supuesto un esfuerzo muy importante de desarrollo, ¿en qué grado se ha producido una mejora?
- Eliminando la referencia ideal correspondiente a la entonación a natural, ¿en qué grado es aceptable el modelo obtenido?

Para obtener respuesta a la primera pregunta, se han planteado hasta ahora dos tipos básicos de metodología: los métodos objetivos, que utilizan alguna medida de distancia entre las curvas de F0 sintéticas y naturales; y los métodos subjetivos, basados en la obtención de opiniones de oyentes humanos [2]. Aunque los primeros pueden parecer más fiables (por su validez matemática), los segundos proporcionan una evaluación del grado de naturalidad basada en la 'percepción'. Sin descartar el desarrollo futuro de una evaluación objetiva de todos los modelos desarrollados, en este trabajo se presentan los resultados obtenidos con un método subjetivo basado en la opinión comparativa entre entonación natural y sintética de un cierto número de individuos. Esta comparación entre entonación natural y sintética se ha llevado a cabo en la prueba I.

Con respecto a la segunda pregunta planteada, está claro que únicamente es posible utilizar pruebas de opinión y se ha realizado una comparación directa entre las entonaciones obtenidas con el modelo de Picos y Valles y el de Fujisaki [3], que ha sido previamente adaptado al euskera [4] Esta comparación de diferentes modelos de entonación sintética se ha realizado en la prueba II.

Finalmente, en un último bloque de pruebas (prueba III) que trata de responder a la tercera pregunta, se han evaluado las entonaciones sintéticas obtenidas con los dos modelos de entonación disponibles. Las características de las pruebas realizadas se explican con más detalle en el siguiente apartado.

## 3. PROCESO DE EVALUACIÓN

### 3.1. Generación de estímulos

Para el desarrollo de la prueba I, se utilizaron un conjunto de 15 frases, cuidadosamente seleccionadas de la base de datos utilizada para la generación del modelo de entonación basado en Fujisaki. Para poder comparar exclusivamente la bondad del modelo de entonación, se sustituyó la frecuencia fundamental de las frases

naturales por la curva de entonación sintética obtenida con el modelo, utilizando un Vocoder LPC. Las frases naturales se reprodujeron también con transcodificación LPC, con el fin de que la calidad de ambas señales fuera similar.

Otro conjunto de 10 frases de similares características también procedentes de la misma base de datos se utilizó para la prueba II, esta vez utilizando el sistema de conversión de texto a voz AhoTTS [5], con ambos modelos de entonación.

Finalmente, para realizar la prueba III, se seleccionó un conjunto de 4 textos cortos (unas 25 palabras) de diferentes tipos: mensajes de correo electrónico y párrafos extraídos de periódicos. Para obtener los estímulos, los textos se sintetizaron con el sistema AhoTTS, utilizando los dos modelos de entonación.

### 3.2. Método de evaluación

Para facilitar la realización de la prueba, se ha desarrollado una aplicación en la que el propio usuario puede controlar la reproducción de los estímulos, e introducir la puntuación correspondiente a cada pareja de estímulos. La prueba se realizó en ambiente de laboratorio, con auriculares. La duración total aproximada de una prueba es de 10 minutos.

Para la prueba I, el usuario escucha los estímulos correspondientes a las entonaciones natural y sintética, y puntúa en una escala de 1 a 5 el grado de similitud entre ambos estímulos (1: ninguna similitud, 5: total similitud). El usuario puede reproducir los estímulos tantas veces como desee. El aspecto de la ventana del usuario para esta prueba se muestra en la figura 1.

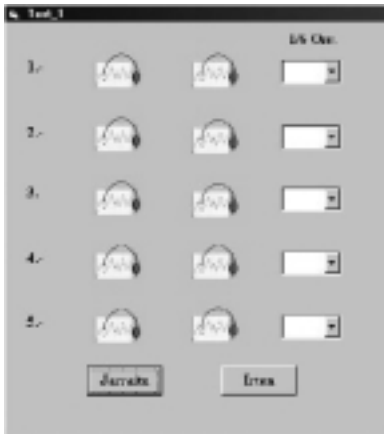

Figura 1. *Aspecto del programa de evaluación para la prueba I.*

Esta prueba incluye tres parejas de estímulos idénticos, cuyo fin es detectar evaluadores no capaces de juzgar adecuadamente la similitud de dos entonaciones.

En el desarrollo de la prueba II, el usuario elige el modelo preferido tras escuchar estímulos correspondientes a ambos modelos de entonación, que se presentan en cualquier orden. Finalmente, durante el desarrollo de la prueba III, el usuario puntúa también en escala de 1 a 5, el grado de satisfacción que le produce el modelo (1: nada satisfecho, 5: muy satisfecho) escuchando de forma independiente los textos sintetizados para cada uno de los modelos, siéndole presentados los estímulos correspondientes a uno y otro modelo en orden aleatorio.

## 4. RESULTADOS

Los resultados provisionales de las pruebas de evaluación son alentadores. En la tabla 1 se muestra el resultado de la prueba I, es decir, la puntuación media con que los evaluadores han calificado el parecido entre las curvas de entonación naturales y las sintéticas obtenidas conforme al modelo de Fujisaki. Se ve que las curvas han sido juzgadas como razonablemente parecidas.

Tabla 1. *Resultado de la prueba I.*

| | Media |
|---|---|
| Máximo | 4.4 |
| Mínimo | 2.4 |
| Valor medio | 3.43 |

En cuanto a la comparación entre las entonaciones sintéticas creadas con el modelo de picos y valles y el de Fujisaki, el 97% de los evaluadores ha preferido éste último.

Finalmente, el resultado de la prueba III se muestra en la tabla 2, en la que puede observarse que el modelo de Fujisaki es evaluado más positivamente que el anterior, siendo la puntuación media alcanzada superior al valor 3 que indica que el modelo es aceptable. El modelo de picos y valles sin embargo no alcanza esta cifra, siendo calificado globalmente como no aceptable.

Tabla 2. *Resultado de la prueba III.*

| | Picos y valles | Fujisaki |
|---|---|---|
| Máximo | 4 | 4 |
| Mínimo | 1 | 3 |
| Valor medio | 2.19 | 3.54 |

## 5. AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por MCYT, TIC2000-1005-C03-03 y TIC2000-1669-C04-03.

## 6. REFERENCIAS

[1] Hernáez, I.; Olabe, J.C; Cuesta, A.; Gandarias R., Etxebarria, P.. "Improving Naturalness in a Text-to-Speech conversión system for the Basque Language". Proc. of the 7th Mediterranean Electrotechnical Conference, Vol. I, pp.61-64, 1994.

[2] Hirst, D., Rilliard, A., Aubergé, V.. "Comparison of subjective and objective evaluation metric for prosody in text-to-speech synthesis". *In Proceedings of 3rd ESCA/COCOSDA Workshop on Speech Síntesis.* 1998.

[3] Fujisaki, H., Hirose, K.. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese". *Journal of Acoustic Society. Jpn*. (E) 5, 4. 1984.

[4] Navas, E.; Hernáez, I.; Armenta, A.; Etxebarria, B.; Salaberria, J.. "Modelling Basque intonation using Fujisaki's model and CARTs". *State of the art in speech synthesis digest, 3/1-3/6*. 2000.

[5] Hernáez, I., Navas, E., Murugarren, J.L., Etxebarria, B.. "Description of the Hatos System for Basque Language", *In Proceedings of 4th ISCA Tutorial & Research Workshop on Speech Síntesis*. 2001.