

PROCEDIMIENTO PARA LA MEDIDA Y LA MODIFICACIÓN DEL JITTER Y DEL SHIMMER APLICADO A LA SÍNTESIS DEL HABLA EXPRESIVA

Carlos Monzo, Ignasi Iriundo y Elisa Martínez

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

{cmonzo, iriundo, elisa}@salle.url.edu

RESUMEN

En este trabajo se presenta un nuevo procedimiento para la medida de los parámetros de calidad de voz (VoQ), el *jitter* y el *shimmer*. Este nuevo procedimiento tiene en consideración la prosodia del enunciado, de manera que su efecto se atenúa antes de realizar la medida de cada uno de los parámetros. El objetivo, además de realizar la medida de una forma más fiable, es el de modificar estos parámetros de forma que puedan ser utilizados en síntesis del habla expresiva, por ello, en paralelo a esta nuevo procedimiento de análisis, se presenta cómo llevar a cabo la modificación de ambos. Finalmente, se realiza una evaluación mediante una prueba perceptual CMOS sobre cuatro estilos expresivos: agresivo, alegre, sensual y triste; provenientes de la salida de un sistema de conversión de texto en habla con modelado prosódico, de modo que se hace un estudio de la utilidad de estos parámetros bajo diferentes situaciones.

1. INTRODUCCIÓN

El reconocimiento automático del habla y la conversión de texto en habla (CTH) son áreas de investigación en las que el habla expresiva se está usando con el objetivo de mejorar la naturalidad de la interacción persona-máquina. Ejemplos de esta investigación los encontramos en estudios sobre reconocimiento de emociones [1] o transformación de voz [2][3]. La prosodia y la calidad de voz (de aquí en adelante VoQ) son parámetros utilizados en la representación del contenido emocional del habla tal y como se presenta en [1][4]. A pesar de que la VoQ ha sido menos estudiada que la prosodia, trabajos recientes proponen ambas informaciones para mejorar el modelado acústico del habla expresiva [5][6].

En este trabajo nos interesa la VoQ, que tradicionalmente ha sido analizada de manera independiente a la prosodia, ya sea en aplicaciones de

patologías de voz [7], donde no se considera por la naturaleza y condiciones de medida, o bien en estudios sobre estilos expresivos [5][6], donde ya se observa que la prosodia se debe considerar, tal y como señala [8].

El objetivo principal de este trabajo, es el de presentar un nuevo procedimiento de análisis y modificación de los parámetros de VoQ, el *jitter* y el *shimmer*, teniendo en cuenta el efecto de la prosodia. En la bibliografía [9][10][11] se muestra como su medida se realiza sin considerar la variación prosódica debida a la expresividad, o emoción transmitida, de forma que es necesario en aplicaciones como síntesis del habla expresiva o reconocimiento de emociones, considerar su efecto para así tratar de cancelarlo y obtener una medida de VoQ sin interferencias. Complementariamente, se evalúa el efecto de añadir el *jitter* y el *shimmer* al habla generada por un CTH basado únicamente en modificación prosódica.

Este artículo está organizado como sigue. En el apartado 2 se introduce el material de voz usado en el diseño y evaluación. El apartado 3 explica el nuevo procedimiento de análisis para el *jitter* y el *shimmer*. Los apartados 4 y 5 presentan la modificación de estos parámetros, así como la evaluación y discusión de los resultados. Para terminar, el apartado 6 muestra las conclusiones alcanzadas.

2. MATERIAL DE VOZ

El material de voz usado para realizar los experimentos sobre los nuevos procedimientos de medida del *jitter* y del *shimmer*, es el mismo que utiliza el CTH desarrollado por el GPM [12], se trata de cinco corpus de habla expresiva (o emocionada): neutro, agresivo, alegre, sensual y triste; en español y grabados por una locutora profesional. En [6] se encuentra una explicación detallada de ellos.

Los procedimientos de análisis y modificación del *jitter* y del *shimmer* se han aplicado sobre muestras de habla sintetizada, generadas a partir del corpus de habla neutra y con el modelo prosódico de la expresividad deseada [13].

3. NUEVA METODOLOGÍA DE ANÁLISIS

Este apartado expone un nuevo procedimiento para la medida del *jitter* y del *shimmer*. Primero se presenta una descripción de cada uno de ellos y posteriormente se explica la propuesta. El cálculo habitual de estos parámetros, como el realizado por la herramienta Praat [10], no tiene en cuenta variaciones ni de F_0 ni de la energía debidas principalmente al efecto de la prosodia.

3.1. Jitter

Según [5], el *jitter* se corresponde a las variaciones de F_0 que existen en el tramo de habla analizado, representadas como un ruido por modulación en frecuencia.

El procedimiento que se propone, parte de la información de marcas de F_0 calculadas únicamente en las zonas sonoras de la señal de voz, usando para ello el algoritmo RAPT [14]. A partir de ellas se calcula la curva de F_0 en cada una de las zonas sonoras y se lleva a cabo una transformación logarítmica utilizando semitonos [15], consiguiéndose así una normalización relativa al tono medio y una mejor representación de la percepción subjetiva de las variaciones de tono. La transformación de hercios a semitonos y su inversa se muestra en las ecuaciones (1) y (2) respectivamente, donde 'ref' es la frecuencia de referencia. Para este trabajo, se ha tomado como referencia la F_0 media del locutor para la expresividad deseada, calculada a partir del correspondiente corpus de síntesis.

$$\text{Hz} = 2^{\text{St}/12} \cdot \text{ref} \quad (1)$$

$$\text{St} = 12 \cdot [\ln(\text{Hz}/\text{ref})/\ln 2] \quad (2)$$

A partir de los valores transformados en semitonos, se realiza una detección de los tramos de crecimiento y decrecimiento del contorno de F_0 a partir de un análisis de la pendiente. Para cada tramo obtenido se lleva a cabo una regresión lineal, que se resta de la curva inicial de F_0 , con el fin de tratar de anular el efecto de la prosodia (véase ejemplo en la Figura 1).

Para terminar, se calcula la variación de F_0 (ΔF_0) entre periodos consecutivos (F_0) tal y como muestra la ecuación (3) y, finalmente, se calcula el valor del *jitter* para cada una de las tramas según la ecuación (4):

$$\Delta F_0_i(j) = F_0_i(j+1) - F_0_i(j) \quad (3)$$

$$\text{jitter}_i = \frac{1}{N} \cdot \sum_{j=1}^N \Delta F_0_i(j)^2 \quad (4)$$

Donde $j = 1:(\text{núm de marcas de } F_0 \text{ en la trama}) - 1$, $i = \text{trama bajo análisis y } N = \text{longitud de } \Delta F_0_i$.

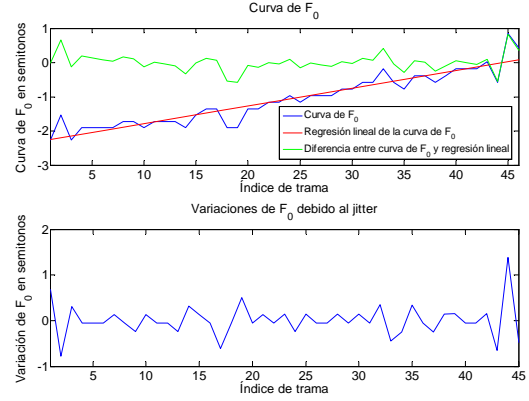


Figura 1. Extracción de la variabilidad de F_0

3.2. Shimmer

El *shimmer* computa las variaciones de amplitud de la forma de onda tal y como se presenta en [5]. Describe un ruido por modulación en amplitud.

El nuevo procedimiento propuesto está inspirado en el expuesto para el *jitter*, por tanto parte de las marcas de F_0 de las zonas sonoras. Se calcula la curva de amplitudes pico a pico máximas, por periodo de F_0 , en cada una de las zonas sonoras, llevando a cabo por último una transformación logarítmica, aplicando el logaritmo natural.

Una vez se dispone de los valores transformados, se elimina el efecto prosódico de la energía igual que se hizo para la F_0 en el apartado 3.1 (véase la Figura 2).

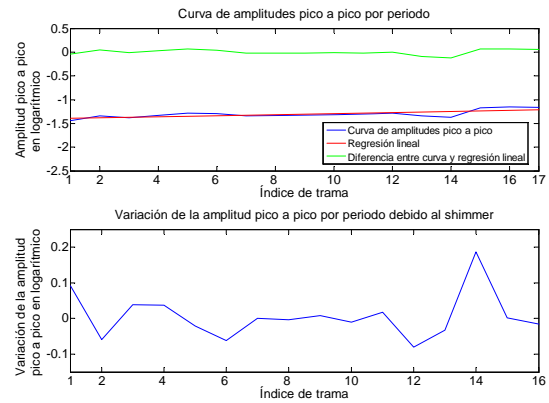


Figura 2. Extracción de la variabilidad de amplitud

En (5) se calcula la variación de amplitud pico a pico (Δcpap) en periodos consecutivos (cpap), presentando en (6) el cálculo de *shimmer* por trama:

$$\Delta \text{cpap}_i(j) = \text{cpap}_i(j+1) - \text{cpap}_i(j) \quad (5)$$

$$\text{shimmer}_i = \frac{1}{N} \cdot \sum_{j=1}^N \Delta \text{cpap}_i(j)^2 \quad (6)$$

Donde $j = 1:(\text{núm de periodos de } F_0 \text{ en la trama}) - 1$, $i = \text{trama bajo análisis y } N = \text{longitud de } \Delta \text{cpap}_i$.

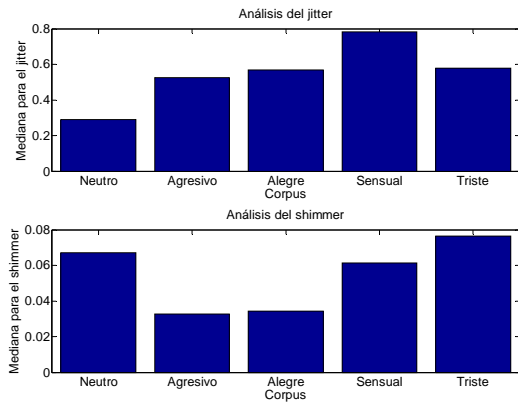


Figura 3. Análisis del jitter y del shimmer sobre los 5 corpus expresivos

4. EXPERIMENTOS

En este apartado se muestra como, a partir del procedimiento de análisis presentado para la medida del jitter y del shimmer, estos parámetros pueden ser modificados. A partir de aquí se evalúa cómo afectan éstos a la identificación de expresividades generadas por un CTH, por un lado, únicamente con modelado prosódico [13], y por otro con modificación de VoQ.

4.1. Modificación del jitter y del shimmer

La modificación para ambos sigue el mismo procedimiento, basándose en la inserción de ruido blanco sobre una curva de F_0 o de amplitudes pico a pico limpias de jitter o de shimmer. Para ello, a partir de la prosodia se calcula la regresión lineal de aquellos tramos donde la curva de interés mantenga su tendencia, y sobre ella se realizará la adición del ruido blanco.

El ruido blanco que se añade tiene como potencia el valor esperado del jitter y del shimmer adecuado a la expresividad que se desea simular. Estos valores se conocen a partir del proceso de análisis sobre los corpus expresivos, presentados en el apartado 2, correspondiéndose al valor de mediana obtenido a partir de estadística descriptiva sobre cada uno de los corpus.

Los resultados obtenidos para los diferentes parámetros jitter y shimmer, usando el procedimiento expuesto, se muestran en la Figura 3. El utilizar la mediana y no la media, se debe a que de este modo, después de analizar las distintas distribuciones, se ha visto que se evitan valores atípicos que puedan desviar el valor medio de la medida.

4.2. Evaluación

Dado el nuevo procedimiento de análisis y modificación del jitter y del shimmer, el siguiente paso ha sido evaluar cómo pueden contribuir, cada uno por separado o bien de forma conjunta, a la síntesis del habla expresiva.

La prueba parte de 5 enunciados sintetizados usando el CTH presentado en [12] con 4 expresividades diferentes: agresiva, alegre, sensual y triste; a partir de aplicar modelos prosódicos para cada una de ellas [13] sobre un enunciado originariamente neutro. Una vez generadas se les aplicará la modificación del jitter y del shimmer para su posterior evaluación. Ésta se realiza mediante una prueba perceptual CMOS [16] usando la interfaz web presentada en [17]. Si se desea profundizar sobre la expresividad obtenida usando únicamente modelado prosódico se recomienda la lectura de [18].

Los enunciados se muestran en parejas, comparando la original sintetizada usando modelos prosódicos con la modificada usando el jitter, el shimmer o ambos parámetros, dando lugar a 60 comparaciones. Cada uno de los evaluadores, 13 en total, eligió si la intensidad de la expresividad presentada en el ejemplo era “mucho más”, “más”, “poco más” o “igual” que la del otro, con una puntuación de: 3, 2, 1, 0, -1, -2 y -3.

5. RESULTADOS Y DISCUSIÓN

En cuanto a los resultados obtenidos, los valores positivos se han reservado para aquellos casos donde, por usar VoQ, la expresividad se muestra con mayor intensidad, mientras que los negativos indican que es la original con solo modelado prosódico (véase la Figura 4).

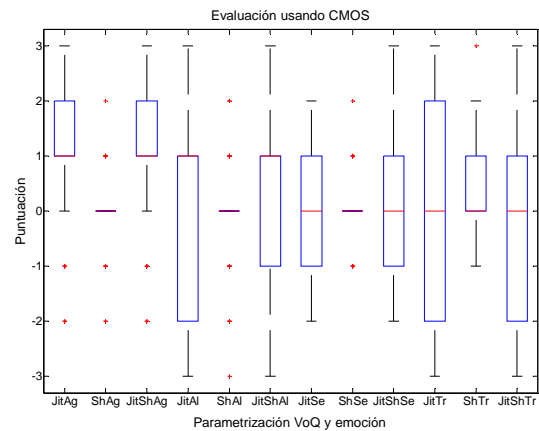


Figura 4. Resultados de la prueba CMOS sobre la VoQ

Por otro lado, el valor de CMOS medido por configuración, calculado como su valor medio, se presenta en la Tabla 1, señalando en cursiva a aquellos donde la VoQ mejora la percepción de la expresividad y en negrita cuando su efecto es más representativo.

Tabla 1. CMOS para 3 configuraciones y 4 expresividades

	Agresiva	Alegre	Sensual	Triste
Jitter	<i>1.06</i>	0.00	-0.12	-0.06
Shimmer	<i>0.12</i>	-0.06	<i>0.08</i>	<i>0.29</i>
Sh + Jit	<i>1.14</i>	<i>0.18</i>	-0.03	-0.31

Como se puede observar en la Figura 4, la expresividad para la que el efecto de la VoQ intensifica en mayor grado su percepción es la “Agresiva”, con un valor superior a 1 en la escala CMOS (véase la Tabla 1). Por otro lado, se tiene que la “Alegre” presenta buenos resultados, mediana igual a 1, a pesar de su dispersión. El resto de expresividades, la “Sensual” y la “Triste”, dan resultados con una elevada dispersión, hecho que hace pensar en su utilidad solamente en ciertos casos. Estos resultados son interesantes en tanto que la “Agresiva” y la “Alegre” daban los peores resultados en estudios que usaban solamente prosodia [18], por tanto, los parámetros *jitter* y *shimmer*, la complementará al generar estas expresividades.

Por último, en cuanto al parámetro de VoQ que contribuye al incremento en la intensidad de la expresividad presentada, es el *jitter* para el caso de la “Agresiva”, el *shimmer* para la “Triste” y una combinación de ambos para la “Alegre” y la “Sensual”.

6. CONCLUSIONES

En este trabajo se ha presentado un nuevo procedimiento para la medida de los parámetros de VoQ, el *jitter* y el *shimmer*. Esta metodología trata de evitar la dependencia con la prosodia de forma que puede ser utilizada en el análisis del habla expresiva.

Visto el procedimiento de análisis, se ha presentado cómo realizar la modificación de estos parámetros, mostrando su utilidad en síntesis del habla expresiva.

Por último, con el objetivo de evaluar la utilidad y dependencia de cada parámetro con una expresividad diferente, se ha llevado a cabo una prueba perceptual CMOS, donde la utilidad de los parámetros de VoQ en la síntesis del habla expresiva ha quedado justificada.

De los resultados obtenidos, se plantea como trabajo futuro, realizar un análisis de la dependencia del lugar del enunciado donde el evaluador centra su atención, pudiéndose así aplicar las modificaciones de VoQ de forma más específica.

7. BIBLIOGRAFÍA

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, y J. Taylor. “Emotion recognition in human-computer interaction”, *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80, 2001.

[2] C. Drioli, G. Tisato, P. Cosi, y F. Tesser, “Emotions and voice quality: experiments with sinusoidal modeling”, *VOQUAL'03*, pp. 127-132, Geneva, 2003.

[3] O. Turk, M. Schröder, B. Bozkurt, y L.M. Arslan, “Voice quality interpolation for emotional text-to-speech synthesis”, *INTERSPEECH*, pp. 797-800, Lisbon, 2005.

[4] C. Gobl, y A. Ní Chasaide, “The role of voice quality in communicating emotion, mood and attitude”. *Speech Communication*, 40, 189-212. 2003.

[5] C. Monzo, F. Alías, I. Iriondo, X. Gonzalvo, y S. Planet, “Discriminating Expressive Speech Styles by Voice Quality Parameterization”, *ICPhS*, pp. 2081-2084, Saarbrücken, 2007.

[6] I. Iriondo, S. Planet, J.C. Socoró, F. Alías, C. Monzo, y E. Martínez, “Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality”, *ICPhS*, pp. 2125-2128, Saarbrücken, 2007.

[7] F. Núñez, P. Corte, C. Suárez, B. Señaris, y G. Sequeiros, “Evaluación perceptual de la disfonía: correlación con los parámetros acústicos y fiabilidad”, *Acta otorrinolaringológica española: Órgano Oficial de la Sociedad Española de Otorrinolaringología y Patología Cérvico-Facial*, vol. 55, no. 6, pp. 282-287, 2004.

[8] M. Swerts, y R. Veldhuis, “The effect of speech melody on voice quality”, *Speech Communication*, vol. 33, pp. 297-303, 2001.

[9] R.E. Slyh, W.T. Nelson, y E.G. Hansen, “Analysis of mrate, shimmer, jitter, and F₀ contour features across stress and speaking style in the SUSAS database”, *ICASSP '99*, vol. 4, pp. 2091-2094, Phoenix, 1999.

[10] P. Boersma, “Praat, a system for doing phonetics by computer”, *Glott International*, vol. 5, no. 9-10, pp. 341-345, 2001.

[11] A. Verma, y A. Kumar, “Introducing Roughness in Individuality Transformation through Jitter Modeling and Modification”, *ICASSP'05*, vol. 1, pp. 5- 8, ISSN: 1520-6149 ISBN: 0-7803-8874-7, Philadelphia, 2005.

[12] F. Alías, e I. Iriondo, “La evolución de la Síntesis del Habla en Ingeniería la Salle”, *2JTH02*, Granada, 2002.

[13] I. Iriondo, J.C. Socoró, L. Formiga, X. Gonzalvo, F. Alías, y P. Miralles, “Modelado y estimación de la prosodia mediante Razonamiento Basado en Casos”, *4JTH06*, pp. 183-188, ISBN 84-96214-82-6, Zaragoza, 2006.

[14] D. Talkin, “A Robust Algorithm for Pitch Tracking (RAPT)”, *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 14, pp. 495-518. Elsevier Science, Amsterdam, 1995.

[15] G. Fant, A. Krucenberg, K. Gustafson, y J. Liljencrants, “A new approach to intonation analysis and synthesis of Swedish”, *Proceedings of Fonetik*, pp. 161-64, Stockholm, 2002.

[16] ITU-P.800, “Methods for subjective determination of transmission quality”, *Recommendation P.800 International Telecommunication Union (ITU)*, 1996.

[17] S. Planet, I. Iriondo, E. Martínez, y J.A. Montero, “TRUE: an online testing platform for multimedia evaluation”, *LREC'08*. Marrakech, 2008.

[18] I. Iriondo, “Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva”, *Tesis Doctoral*, Barcelona, 2008.