(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 763 937**

(51) Int. CI.:

G10L 25/60 (2013.01)
G10L 25/66 (2013.01)
G10L 25/21 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **31.08.2016 E 16186498 (8)**

(97) Fecha y número de publicación de la concesión europea: **09.10.2019 EP 3291234**

(54) Título: **Procedimiento para la valoración de una calidad de un uso de la voz de un hablante**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
01.06.2020

(73) Titular/es:

DIGITHEP GMBH (100.0%)
Oranienburger Straße 66
10117 Berlin, DE

(72) Inventor/es:

HÖNIG, FLORIAN

(74) Agente/Representante:

LINAGE GONZÁLEZ, Rafael

DESCRIPCIÓN

Procedimiento para la valoración de una calidad de un uso de la voz de un hablante

La presente invención se refiere a un procedimiento para la valoración de una calidad de un uso de la voz de un hablante. Además, se refiere a un programa de tratamiento de datos para la valoración automatizada, asistida por ordenador de una calidad de un uso de la voz de un hablante y un equipo de tratamiento de datos asistido por ordenador, que contiene un programa de tratamiento de datos semejante.

En el estado de la técnica se conocen diferentes procedimientos y métodos para la valoración del tipo de una vocal hablada, p. ej. Por GARELLEK, M. "The timing and sequencing of coarticulated non-modal phonation in English and White Hmong", Journal of Phonetics, 2012, volumen 40, páginas 152 a 161, o también por STEVENS, Kenneth N., Acoustic Phonetics, Londres, 1998, páginas 422 a 451.

La valoración de la calidad de un uso de la voz, en particular según un así denominado uso de la voz duro o suave, tiene importancia para el entrenamiento del habla. Esto es aplicable en particular, pero no exclusivamente, para el entrenamiento del habla para tartamudos. Pero para la formación de logopedas, que deben aprender un uso consabido de su voz, a fin de ofrecer aún más esta capacidad, tiene importancia la valoración de la calidad de un uso de la voz, como también p. ej. en la formación de personas que trabajan mucho con su habla.

El tartamudeo es un trastorno del habla, en el que se producen repeticiones, alargamientos y bloqueos involuntarios. En el mundo está afectada aprox. el 1% de la población. Los tartamudos padecen en parte considerablemente miedos sociales y exclusión. El trastorno del habla se considera como no curable según el nivel del conocimiento actual, pero a través de ejercicios de pronunciación se puede dominar bastante bien y contenerse en sus repercusiones. Así, por un lado, existe la necesidad de desarrollar y ofrecer programas de ejercicios de pronunciación apropiados; desde el punto de vista comercial existe un mercado no insignificante para las prestaciones de servicios o productos correspondientes.

Un enfoque difundido y eficaz de un modo de proceder correspondiente es el enfoque técnico del habla. En este caso se aprende un nuevo modo de hablar modificado, en el que los eventos de tartamudeo aparecen raras veces. Una de las variantes más prominentes de ello es Websters "Precision Fluency Shaping Program" (PFSP), que ya se ha desarrollado al comienzo de los años 70 y se destaca entre otros por usos de la voz suaves, alargamiento de sonidos, debilitamiento y adherencia discordante de las consonantes.

Un componente del trabajo con los tartamudos es con frecuencia un feedback asistido por ordenador sobre la pronunciación. Una adaptación alemana de un programa PFSP se estable bajo el nombre de terapia de tartamudeo de Kassel (KST) y se reconoce como muy efectiva.

Un aspecto fundamental de la Fluency Shaping (conformación de fluencia) y enfoques afines son los usos de la voz suaves. Gracias a ello no se producen en absoluto bloqueos. Los usos de la voz suaves se usan por los tartamudos con frecuencia ya de forma intuitiva a fin de resolver los bloqueos. En el caso de usos de la voz suaves, la generación de tonos comienza por la glotis, sin que anteriormente se haya formado un cierre por la glotis. De este modo la voz empieza de forma suave y baja. Por el contrario se destaca un uso de la voz duro mediante un cierre anterior de la glotis, que resulta entonces en forma de un ruido oclusal. Estas diferencias se pueden comprender, por ejemplo, mediante la comparación de una electroglotografía y señal vocal o en vídeos a alta velocidad de la glotis. Los usos de la voz suaves se destacan además por la ausencia de laringizaciones (voz chirriante, es decir, oscilaciones de las cuerdas vocales irregulares y de baja frecuencia).

La Websters Fluency Shaping y métodos afines, como KST, usan un análisis automático asistido por ordenador de la señal vocal, a fin de darle un feedback al practicante respecto a sus usos de la voz. En los métodos existentes para el reconocimiento de los usos de la voz y la valoración de una calidad de los mismos se utiliza que la suavidad del uso de la voz se manifiesta habitualmente en un aumento progresivo de la sonoridad.

Por ejemplo, por Webster se usa el desarrollo de la sonoridad durante los primeros 100 milisegundos del uso de la voz, a fin de llegar a una clasificación automática de un uso de la voz suave o dura. Otros autores también describen el uso del desarrollo de la sonoridad de una locución para valorar automáticamente los usos de la voz. Las descripciones correspondientes se encuentran p. ej. en el documento 4,020,567.

Bajo una serie de parámetros, que describen el desarrollo de la sonoridad se identifica como el mejor el logaritmo del así denominado Rise Time (así designado H.F. Peters et al in "Perceptual judgement of abruptness of voice onset in vowels as a function of the amplitude envelope", Journal of Speech and Hearing Disorders, vol. 51, nº 4, pág. 299-308, 1986 el tiempo que transcurre entre el alcance del 10% y 90% de la sonoridad máxima). Bajo las condiciones muy

controladas de este estudio, este parámetro se correlaciona muy bien con la estimación perceptiva (gradual) de la suavidad de la voz.

En el documento US 2012/0116772 A1 se describe un sistema de servidor de cliente general para favorecer una logopedia general. En un terminal móvil se graba el habla del paciente y se analiza automáticamente. La señal vocal se le transmite a un servidor, a fin de efectuar otros análisis automáticos y le ofrecen al terapeuta posibilidades para intervenir en la terapia. Los análisis automáticos incluyen también el análisis del uso de la voz. Asimismo se menciona la Fluency Shaping como enfoque de terapia para el tartamudeo. Sin embargo, la extracción de parámetros acústicos para el análisis automático se describe de forma muy amplia y genérica. Para el análisis de los usos de la voz no se expone más en detalle cómo y mediante qué parámetros se realiza esto.

En tanto que se describe en detalle, los sistemas conocidos por el estado de la técnica para la clasificación del uso de la voz solo recurren al desarrollo de la sonoridad. Esto es problemático desde dos puntos de vista:

en primer lugar, un aumento de la sonoridad baja, paulatina no es suficiente ni necesario para un uso de la voz suave. Por ejemplo, un uso de la voz chirriante puede presentar un aumento de la sonoridad semejante. No obstante, en la Fluency Shaping se debe evitar el chirrido en cualquier caso.

A la inversa un uso de la voz puede ser suave y, sin embargo, aumentar rápidamente en la sonoridad. Así, la sonoridad se ve como parámetro insuficiente o incompleto para la determinación y clasificación de un uso de la voz como “suave” o justamente “no suave”.

Esto es congruente con la ayuda habitual de producir usos de la voz bajos, que se vuelven paulatinamente más altos , pero esto es problemático a pesar de todo. Bajo condiciones muy controladas los parámetros pueden conducir al éxito a partir del desarrollo de la sonoridad. Bajo condiciones realistas, es decir, usos de la voz no prototípicos y con accesorios de audio heterogéneos y/o económicos y por consiguiente menos exactos es claramente más difícil en cambio la tarea de la clasificación y no se puede resolver de forma suficientemente exacta por sí solo a través del análisis del desarrollo de la sonoridad.

Ante estos antecedentes, el objetivo de la invención es especificar un procedimiento para la valoración de la calidad de un uso de la voz, con el que se puedan reconocer de forma fiable los usos de la voz suaves de hablantes muy diferentes en un modo de proceder automatizado, en particular asistido por ordenador.

Este objetivo se consigue mediante un procedimiento para la valoración de una calidad de un uso de la voz de un hablante con las características de la reivindicación 1. Perfeccionamientos ventajosos del procedimiento según la invención están designados en las reivindicaciones dependientes 2 a 5. Otro aspecto de la solución consiste además en un programa de tratamiento de datos con las características de la reivindicación 6, respecto al que están especificados perfeccionamientos ventajosas en las reivindicaciones 7 a 9. Finalmente con la invención también se especifican un equipo de tratamiento de datos asistido por ordenador con las características de la reivindicación 10 y variantes de configuración ventajosas respecto a este con las características opcionales de las reivindicaciones 11 y 12.

En el procedimiento según la invención para la valoración de una calidad de un uso de la voz de un hablante se detecta por lo tanto una señal vocal acústica del hablante y se convierte en una señal vocal digital. La señal vocal digital se analiza entonces en su desarrollo temporal, para

a. determinar en el desarrollo temporal de la señal vocal digital un momento del uso de la voz del hablante,
b. determinar una frecuencia fundamental de la señal vocal en el momento del uso de la voz,
c. determinar a partir de la señal vocal digital en un intervalo de tiempo predeterminado desde el momento del uso de la voz de forma resuelta temporalmente el desarrollo de la energía contenida en la señal vocal a la frecuencia fundamental;
d. determinar a partir de la señal vocal digital en el intervalo de tiempo predeterminado de forma resuelta temporalmente el desarrollo de la energía contenida en la señal vocal en al menos un múltiplo armónico de la frecuencia fundamental;
e. determinar el desarrollo temporal de la relación de las energías determinadas en las etapas c y d.

Recurriendo a las energía determinadas bajo c. y d. y su relación determinada según e. se concluye un uso de la voz suave cuando, en el intervalo de tiempo, la relación de las energías determinada en e. anterior se domina en primer lugar por la energía en la frecuencia fundamental y solo en el desarrollo posterior del intervalo de tiempo predeterminado en el lapso de tiempo Δt se desplaza la relación de las energías en beneficio de la(s) energía(s) en el/los múltiplo(s) armónico(s) de la frecuencia fundamental.

Aquí, así en otras palabras no se tienen en cuenta la sonoridad y su desarrollo (correspondientemente la energía) de todo el evento de habla en y después del uso de la voz, sino que se realiza una observación detallada según distintas fracciones de la señal de la voz.

Los datos grabados y evaluados por los inventores muestran de hecho que el desarrollo de sonoridad no conduce a una fiabilidad satisfactoria de la clasificación del comportamiento objetivo de un uso de la voz suave.

El procedimiento según la invención utiliza el conocimiento adquirido ahora por los inventores de que los usos de la voz suaves se destacan ante todo por una calidad de la voz especial al comienzo del evento de habla. Dado que la amplitud de la oscilación de las cuerdas vocales se aumenta solo lentamente en el caso de un uso de la voz suave, tal y como los inventores han reconocido, en la señal vocal está representada en primer lugar principalmente la frecuencia fundamental, en cambio, apenas están presentes sus múltiplos armónicos. La señal vocal es aproximadamente sinusoidal con el uso de la voz y en la primera fase subsiguiente, según se muestra esto en la fig. 1. Solo si las oscilaciones de las cuerdas vocales han alcanzado su máximo, existe un cierre periódico de las cuerdas vocales y por consiguiente la producción de la voz normal con sus fracciones fuertes de múltiplos armónicos.

En el caso del uso de la voz duro, en cambio, la solución del cierre de glotis anterior inicia directamente la vocalización normal con cierre de glotis periódico y la presencia de múltiplos armónicos de la frecuencia fundamental. Esto se muestra en la fig. 2. Este aspecto de la calidad de la voz se extrae con ayuda de parámetros acústicos apropiados, según la invención a través de la relación de energía de las fracciones de oscilaciones a la frecuencia fundamental respecto a las fracciones de oscilaciones en los múltiplos armónicos en un intervalo de tiempo predeterminado. P. ej. este aspecto se puede valorar mediante la relación de las energías del primer armónico (es decir, la frecuencia fundamental) y el segundo múltiplo armónico de la frecuencia fundamental medido en los primeros 10 milisegundos después del uso de la voz. Así, en el marco de la invención no se requiere forzosamente observar las energía de todos los múltiplos armónicos de la frecuencia fundamental. Dado que habitualmente los armónicos de orden más bajo vibran con bastante más energía que los armónicos de mayor orden, puede ser suficiente concentrarse solo en los armónicos de orden más bajo, p. ej. justamente en el segundo armónico (doble de la frecuencia fundamental).

En el marco de la invención, como energía de la frecuencia fundamental Fo también se puede adoptar la suma de las energías en el rango de $0{,}5^{*}F_o$ a $1{,}5^{*}F_o$, para considerar así una “difuminación” de la frecuencia y de la energía contenida en esta. La energía así determinada con la frecuencia fundamental se puede poner entonces p. ej. en relación con la banda de frecuencia ancha, situada por encima de la frecuencia fundamental (p. ej. la energía en el rango entre 50 y 4000 Hz), a fin de caracterizar entonces el uso de la voz a partir del desarrollo temporal de esta relación.

Para distintos ejemplos se produjeron los siguientes valores de estas relaciones de energía en los ensayos:

a. relación frecuencia explorada en el segundo armónico de la frecuencia fundamental (decibelios; promediado sobre los primeros 50 ms):

- hablante masculino, prototípicamente suave: p. ej. -19,1 o -13,6; prototípicamente duro p. ej. 1,6 o 3,7
- hablante femenina, prototípicamente suave: p. ej. -15,8 o -15,9; prototípicamente dura p. ej. 8,9 o 11,2
- muestra anotada practicantes: suave: -13,6 (valor medio) ± 12,1 (desviación estándar); duro: -3,2 (valor medio) ± 15,7 (desviación estándar)

Se puede reconocer que aquí se puede realizar una buena diferenciación entre usos de la voz suaves y duros debido a la relación de energías.

b. relación de energías sumadas de $0{,}5^{*}F_0$ a $1{,}5^{*}F_0$ respecto al resto en el rango de frecuencia 50-4000 Hz (decibelios; promediado sobre los primeros 50 ms):

- hablante masculino, prototípicamente suave: 10,6 o 9,8; prototípicamente duro -6,0 o -7,4
- hablante femenina, prototípicamente suave: 9,3 o 10,9; prototípicamente dura -9,4 o -12,0
- practicante: suave: 7,0 (valor medio) ± 9,9 (desviación estándar); duro: -2,9 (valor medio) ± 12,4 (desviación estándar).

Aquí también se puede reconocer adecuadamente la diferenciación que se puede obtener mediante el modo de proceder según la invención en una observación semejante de las relaciones de energía relevantes.

En el caso del modo de proceder según la invención se deben reconocer en primer lugar los usos de la voz en el desarrollo de señal de la señal vocal digital, es decir, identificarse y localizarse en el desarrollo temporal de la señal. Para ello se puede subdividir la señal vocal ventajosamente en segmentos sonoros y sordos. Dado que la

diferenciación entre sonoro y sordo está sujeta a errores inherentes debido a las propiedades locales de la señal vocal, se ofrece usar un procedimiento que utilice las condiciones de consistencia globales, para llegar a una subdivisión lo más robusta posible. Esto se puede hacer p. ej. por algoritmos para la extracción del desarrollo de la frecuencia fundamental. Aquí se necesita en primer lugar solo el “producto secundario” de la subdivisión en segmentos sonoros / sordos.

En el marco de la invención se usa para ello preferiblemente el algoritmo RAPT según David Talkin (véase D. Talkin, Speech Coding and Syntesis, Elsevier Science, 1995, vol. 495, pág. 495-518, cap. 14 A Robust algorithm for Pitch Tracking (RAPT)), que debido a la falta de múltiplos armónicos en el caso de usos de la voz suaves es más apropiado para la segmentación precisa que p. ej. algoritmos que trabajan en el rango de frecuencia.

Para minimizar las falsas alarmas, la segmentación todavía se puede alisar con ayuda de operadores morfológicos, como se describe p. ej. por H. Niemann en Klassification von Mustern, Verlag Springer 1983, 2ª edición, disponible en http://www5.cs.fau.de/filead min/Persons/NiemannHeinrich/klassifikation-vonmustern/m00-www.pdf. Junto al momento del uso de la voz, la segmentación también indica la duración de la fase asociada.

El intervalo en el que se determina - medido desde el momento del uso de la voz - de forma resuelta temporalmente el desarrollo de las energías a la frecuencia fundamental (= primer armónico) y uno o varios múltiplos armónicos, puede presentar en particular una longitud de 7,5 a 200 milisegundos, preferentemente de 50 a 100 milisegundos.

Se puede concluir un uso de la voz suave, en particular luego cuando el lapso de tiempo Δt, dentro del que se desplaza la relación de las energías en beneficio de la energía en múltiplos armónicos de la frecuencia fundamental, se sitúa entre 50 y 100 milisegundos.

La determinación de la frecuencia fundamental, asimismo como el análisis de los espectros se realiza p. ej. a intervalos de 5 a 20 milisegundos, en particular a intervalos de 10 milisegundos. La determinación regular de la frecuencia fundamental es importante dado que también se puede modificar esta durante el desarrollo del uso de la voz, en general también se modifica, y el procedimiento debe hacer referencia a la frecuencia fundamental correcta para un análisis exacto.

En conjunto este método para la detección de usos de la voz representa un modo de proceder claramente más robusto que métodos en base puramente a la sonoridad, según se usan en los procedimientos existentes. Otra ventaja es que no se necesita una calibración, mientras que en los procedimiento en base a la sonoridad se debe ajustar o estimar siempre un valor umbral.

En el procedimiento según la invención se puede recurrir a un uso de otros parámetros acústicos relevantes y la modelización directa del comportamiento objetivo. Esto puede aumentar aún más la elevada fiabilidad ya alcanzada mediante el modo de proceder arriba descrito.

A este respecto, para utilizar toda la información relevante posible en la señal vocal y lograr por consiguiente la robustez y fiabilidad máxima, se pueden derivar y calcular distintos parámetros acústicos en una pluralidad de variantes de la señal vocal digital.

La reproducción conforme a espacios de parámetros multidimensionales de las clases a reconocer (usos de la voz suaves / duros) se realiza entonces en particular con ayuda de procedimientos impulsados por datos mediante una muestra anotada, típicamente una colección de usos de la voz que se han valorado y clasificado por los expertos. El procedimiento trabaja entonces de forma completamente automática. Otros grupos de parámetros, que se pueden observar, van dirigidos al desarrollo de la sonoridad (que se tiene en cuenta en el estado de la técnica como único parámetro) y parámetros dominados en el espectro, que se dedica en particular a una consideración de las energías en distintas frecuencias y por consiguiente indirectamente también a la calidad de la voz. Un modo de proceder a modo de ejemplo, que también tiene en cuenta parámetros semejantes, se describe además más en detalle abajo.

Con la invención también se proporciona un programa de tratamiento de datos para la valoración automática, asistida por ordenador de una calidad de un uso de la voz de un hablante con

a. un módulo de análisis del uso de la voz, que está establecido para determinar a partir de un desarrollo temporal de una señal vocal digital obtenida a partir de una señal vocal acústica del hablante un momento del uso de la voz del hablante,
b. un módulo de determinación de la frecuencia fundamental, que está establecido para determinar una frecuencia fundamental de la señal vocal en el momento del uso de la voz,
c. un módulo de determinación de la energía de la frecuencia fundamental, que está establecido para determinar a partir de la señal vocal digital en un intervalo de tiempo predeterminado desde el momento del uso de la voz de

forma resuelta temporalmente el desarrollo de la energía contenida en la señal vocal a la frecuencia fundamental;
d. módulo de determinación de la energía de los armónicos que está establecido para determinar a partir de la señal vocal digital en el intervalo de tiempo predeterminado de forma resuelta temporalmente el desarrollo de la energía contenida en la señal vocal en múltiplos armónicos de la frecuencia fundamental;
e. un módulo de determinación de la relación, que está establecido para determinar el desarrollo temporal de la relación de las energías determinadas por el módulo de determinación de la energía de la frecuencia fundamental y el módulo de determinación de la energía de los armónicos y concluir un uso de la voz suave cuando, en el intervalo de tiempo, la relación de las energía se domina en primer lugar por la energía en la frecuencia fundamental, solo durante el desarrollo posterior del intervalo de tiempo predeterminado se desplaza la relación de las energías en beneficio de las energías en los múltiplos armónicos de la frecuencia fundamental.

Con un programa de tratamiento de datos semejante se puede realizar el procedimiento descrito anteriormente en equipos de tratamiento de datos asistidos por ordenador.

El programa de tratamiento de datos puede contener en particular y ventajosamente también un módulo de digitalización para la generación de la señal vocal digital a partir de la señal vocal acústica.

El módulo de uso de la voz puede usar ventajosamente un algoritmo RAPT según David Talkin para la determinación del momento del uso de la voz, según está designado más en detalle anteriormente.

En particular, el programa de tratamiento de datos puede estar configurado como software de aplicación (así denominada App) para un terminal móvil, como en particular un teléfono inteligente o una tablet-PC.

Además, con la invención se especifica un equipo de tratamiento de datos asistido por ordenador, que contiene un programa de tratamiento de datos, según se describe arriba. A este respecto, el programa de tratamiento de datos está instalado en particular listo para funcionar en el equipo de tratamiento de datos.

Ventajosamente el equipo de tratamiento de datos puede presentar una entrada para la recepción de una señal vocal acústica. Una entrada semejante puede ser p. ej. un micrófono. Pero alternativamente también se puede conectar un componente grabador con el equipo de tratamiento de datos, en el que ya se realiza una digitalización de la señal vocal acústica y que le transmite entonces la señal digital en alta resolución al equipo de tratamiento de datos (por cable o de forma inalámbrica, p. ej. a través de una interfaz según el estándar Bluetooth). Un componente grabador semejante puede estar configurado p. ej. mediante el micrófono de un auricular.

El equipo de tratamiento de datos puede estar configurado en particular como terminal móvil, p. ej. como teléfono inteligente o table-PC.

A continuación - también remitiendo a las figuras adjuntas - se describe y explica todavía de nuevo más detalladamente una forma de realización posible de un procedimiento según la invención.

A este respecto muestran en las figuras adjuntas:

Fig. 1 un análisis de la voz en el caso de un uso de la voz suave. A este respecto, arriba con cruces unidas por líneas se muestra el desarrollo extraído con el algoritmo RAPT según David Talkin de la frecuencia fundamental, donde el desarrollo de la frecuencia fundamental indica además el momento del uso de la voz con aprox. t = 1,61 s. La representación central muestra un espectrograma binarizado (píxel ennegrecido, cuando la intensidad de una frecuencia en el momento correspondiente es mayor que un valor umbral, aquí - 25 dB). La frecuencia fundamental y sus múltiplos armónicos se pueden reconocer allí como líneas horizontales paralelas. Se puede reconocer que la línea más inferior, que muestra la frecuencia fundamental, está aislada en primer lugar y solo en el desarrollo posterior (desde aprox. t = 1,68 se acompaña por múltiplos armónicos. Abajo está representada la señal vocal. Al comienzo del uso de la voz con aprox. t = 1,65 seg. existe aproximadamente una oscilación sinusoidal sencilla.

Fig. 2 en una representación comparable a la fig. 1 con las representaciones de desarrollo de la frecuencia fundamental según el algoritmo RAPT (arriba), espectrograma (medio) y señal vocal (abajo) de un análisis de la voz en el caso de un uso de la voz duro. El momento del uso de la voz se sitúa aquí en aprox. t = 5,24 seg. En el espectrograma se puede reconocer que aquí a diferencia del caso mostrado en la fig. 1 del uso de la voz suave los múltiplos armónicos comienzan prácticamente simultáneamente con la frecuencia fundamental. En la señal vocal también se puede reconocer que directamente con el uso comienza la oscilación compleja (no solo sinusoidal).

Fig. 3a-c cada vez una curva de característica operativa del receptor de un procedimiento que trabaja según un sistema automático, controlado por software para distintos subconjuntos de características. (Cuanto más a la izquierda arriba discurre la curva tanto mejor. La primera bisectriz (falsa alarma = cuota de acierto) se corresponde con la línea base

al azar.) La representación en la fig. 3a muestra una primera evaluación puramente según la sonoridad, según se realiza en el estado de la técnica. La representación en la fig. 3b muestra un evaluación limitada al análisis a prever en cualquier caso de la relación de energías de la frecuencia fundamental respecto al múltiplo armónico / múltiplos armónicos. La fig. 3c muestra un análisis según otros criterios como sonoridad y relación de energías de la frecuencia fundamental respecto al múltiplo armónico / múltiplos armónicos y prueba que también otros parámetros pueden tener una influencia (adicional) en la calidad del análisis en la evaluación. Simultáneamente la fig. 3b muestra que está especialmente marcada la influencia del parámetro "relación de energía de la frecuencia fundamental respecto al múltiplo armónico / múltiplos armónicos" y posibilita una valoración especialmente adecuada de la calidad del uso de la voz. AUC (área bajo la curva; 0,5 = línea base al azar, 1 = perfecto): sonoridad (fig. 3a): 0,424; 0,466; 0,525; 0,464; 0,403; Frecuencia fundamental / múltiplo armónico (Fig. 3b): 0,704; 0,688; 0,466; 0,792; 0,530; resto (Fig. 3c): 0,634; 0,736; 0,512; 0,518; 0,552.

Fig. 4 una curva de característica operativa del receptor para el reconocimiento de distintos tipos de errores versus usos suaves en base a una versión del procedimiento según la invención, en el que la comparación de energías de la frecuencia fundamental / múltiplo armónico se ha complementado con la evaluación de otros parámetros para el reconocimiento de otro uso de la voz suave (se trata de la unión de los subconjuntos observados en la fig. 3a-c de las características). AUC: 0,725; 0,780; 0,481; 0,675; 0.573.

Para la realización del procedimiento según la invención se puede proceder en una variante de realización tal y como se describe a continuación.

1. Segmentación

En primer lugar se identifican y localizan todos los usos de la voz de una locución. Para ello se subdivide la señal vocal en segmentos sonoros y sordos. Para ello se usa el algoritmo RAPT según David Talkin. A este respecto se asume que la frecuencia fundamental es de al menos 50 Hz y como máximo 500 Hz. Para minimizar las falsas alarmas, la segmentación todavía se alisa con ayuda de operadores morfológicos (cierre con 0,1 s, apertura con 0,25 s, en último término cierre con 0,5 s). Junto al momento del uso de la voz, la segmentación también indica la duración de la fase asociada.

2. Cálculo de características

2.1. Características base

En primer lugar se extraen las propiedades de la señal vocal, que constituyen posteriormente la base de varios parámetros acústicos. Estos son:

Sonoridad: Para la ventana de análisis a corto plazo (longitud 40 milisegundos, incremento 10 milisegundos, de ventana von-Hann) se calcula la sonoridad instantánea a lo largo del tiempo a partir de la energía (tomada en logaritmo) de la señal. Anteriormente se sustrae el valor medio de la señal a fin de compensar una compensación eventual del hardware de grabación económico y correspondientemente de menor calidad.

Sonoridad pseudo-período-síncrona: Para calcular la sonoridad de la señal vocal lo más localmente posible (independientemente de períodos de oscilación adyacentes, a ser posible esencialmente más bajos o altos), la ventana de análisis se adapta conforme a la frecuencia fundamental extraída - preferentemente con el algoritmo RAPT (duración del período doble, ventana von-Hann). La frecuencia fundamental actual es F_0; entonces la duración de la ventana de análisis es $2/F_0$. Esto se corresponde con un índice de exploración de la señal vocal de F_s (todas las señales vocales se muestrean a F_s = 16 kHz) con el siguiente número de valores de exploración: Valor redondeado de $2F_s/F_0$.

Espectro a corta plazo: El espectro de cuadrados absolutos (tomado el logaritmo) se calcula a partir de ventanas de análisis con 25 ms de anchura y por debajo de 10 ms de incremento y la ventana de von-Hann.

La transformada de Fourier discreta (DFT) se calcula con ayuda de la transformada de Fourier rápida (FFT). Para elevar la resolución de frecuencia se llena con ceros, de modo que se producen 2048 valores de entrada. Se usan bibliotecas de programas estándares para el cálculo.

Espectro a corto plazo pseudo-período-síncrono: Para el mantenimiento de un espectro lo más local posible se adapta la ventana de análisis a la frecuencia fundamental extraída (duración de período cuadruplicada). Duración de la ventana de análisis $4/F_o$, es decir, con el valor redondeado de $4F_s/F_o$ valores de exploración.

Espectro a corto plazo de la correlación cruzada normalizada: Para calcular el espectro de frecuencia, que se

corresponde exactamente con la información, que usa el algoritmo RAPT usado preferentemente para la extracción de la frecuencia fundamental, el espectro de frecuencia se calcula directamente de la correlación cruzada normalizada. A este respecto se modifica un método para el cálculo del espectro a corto plazo de la función de autocorrelación, en tanto que en lugar de la autocorrelación se usa la correlación cruzada normalizada. Así, el espectro de cuadrados absolutos se calcula a partir del valor de la DFT de la correlación cruzada. Esto posibilita un cálculo especialmente preciso de la energía de los distintos múltiplos armónicos de la frecuencia fundamental. Para la correlación cruzada se usa, como en el algoritmo RAPT, la anchura de ventana de 7,5 ms y la anchura total conforme a la frecuencia fundamental asumida mínima (50 Hz). Para elevar la resolución de frecuencia se llena con ceros antes de la DFT, de modo que se producen 2048 valores de entrada.

2.2. Características

A continuación se exponen las 908 características calculadas, usadas para la realización del procedimiento en la variante aquí descrita (parámetros acústicos, que se usan para la clasificación). A este respecto se debe atender a que no necesariamente se necesitan todas las características expuestas para el éxito del procedimiento. Pero según la invención se tiene en cuenta en cualquier caso la característica “frecuencia fundamental / múltiplo armónico”. Pero en una variante del procedimiento - que no está incluida en la invención aquí reivindicada, que representa simultáneamente una invención independiente - se pueden combinar y recurrir a otras características bajo exclusión de la característica de “frecuencia fundamental / múltiplo armónico”.

Se puede recurrir a otras características, pero no se debe. A este respecto es difícil fijar empíricamente cuáles de un conjunto de características dadas son realmente necesarias. Por lo tanto es habitual en un reconocimiento de patrones utilizar muchas características potencialmente relevantes, y confiarse a un procedimiento impulsado por datos, efectuando la ponderación de las características individuales. Esto conduce - supuestos procedimientos de clasificación robustos y modernos - según la invención a un sistema de clasificación más fiable que la selección manual de unos pocos parámetros muy prometedores.

Cuando en primer lugar se habla de intervalos y para ello se hacen indicaciones de tiempo, entonces la indicación de tiempo 0 ms se refiere a un momento del uso de la voz determinado (con el algoritmo RAPT).

Ruido de fondo: el cálculo de la mayoría de las características se influye en una cierta medida por ruidos de fondo base. Para darle la posibilidad al sistema de clasificación de compensarlo se usan las características que describen la sonoridad del ruido de fondo. La sonoridad del ruido de fondo se estima en primer lugar como el mínimo de la sonoridad, se alisa (tras la toma de logaritmo) con una ventana rectangular de 50 ms. Para lograr la independencia del control de un micrófono de grabación, se normaliza con la sonoridad de la voz estimada, se calcula como valor medio antes o después de la toma de logaritmo, o como mediana de la sonoridad en segmentos sonoros. (Así se producen tres ventajas.)

Sonoridad del uso: La sonoridad durante un intervalo al comienzo del uso se usa en distintas variantes como característica. (Longitudes de intervalo: 10 ms, 50 ms, 100 ms o 200 ms; sonoridad: independiente de sonoro / sordo, solo sonoro o pseudo-período-síncrono; tres normalizaciones distintas como arriba; se producen 4x3x3 = 36 características).

Sonoridad del uso directo: Sonoridad en el momento del uso, en distintas variantes (normalizada con los tres sonoridades de la voz estimadas, o la sonoridad durante los primeros 50 ms, 100 ms o 200 ms; se producen seis características).

Amplitud: Modificación de sonoridad durante toda la fase (máximo / mínimo, cuantil 99%/1%, cuantil 95%/5%, cuantil 90%/10%; sonoridad: independiente de sonoro / sordo, solo sonoro o pseudo-período-síncrono 4x3 = 12 características).

Amplitud local: Modificación de sonoridad al inicio del uso (intervalos: 0 a 50 ms, 0 a 100 ms, 0 a 200 ms, 10 a 50 ms, 10 a 100 ms, 10 a 200 ms, 50 a 100 ms, 100 a 200 ms; sonoridad: independiente de sonoro/sonoro, solo sonoro o pseudo-periodo-síncrono; se producen 8x3 = 24 características).

Aumento de la sonoridad: Ascenso de las rectas de regresión al inicio del uso (intervalos e intensidades como arriba; se producen 24 características).

Bajada de la sonoridad: Se acumula si o cuan intensamente baja localmente la sonoridad al inicio del uso (en el caso de ascenso monótono resulta 0. Intervalos: 50 ms, 100 ms, 200 ms; tres sonoridades diferentes como arriba; se producen 3x3=9 características).

Frecuencia fundamental / múltiplo armónico: Para reproducir la diferencia decisiva en la calidad de la voz entre usos suaves y duros se calcula la energía relativa de la frecuencia fundamental en distintas variantes. (Espectro a corto plazo, espectro a corto plazo pseudo-período-síncrono o espectro a corto plazo de correlación cruzada normalizada; cálculo de la energía a partir de un único valor de frecuencia (índice del coeficiente DFT: valor redondeado de F_0 x 2048/F_s) o acumulado (intervalo de frecuencia simétrico, anchura = frecuencia fundamental: se suman los valores absolutos de los coeficientes DFT de índice redondeado0,5 x F_0 x 2048/F_s hasta índice (redondeado) 1,5 x F_0 x 2048/F_s antes de la toma del logaritmo); normalización mediante la energía del 2º múltiplo armónico (variante de cálculo como en la frecuencia fundamental) o mediante energía global en 50 a 4000 Hz; intervalos: 0 a 10 ms, 0 a 50 ms, 0 a 100 ms, 0 a 200 ms, 10 a 50 ms, 10 a 100 ms, 10 a 200 ms, frase completa, 200 ms hasta final de frase; acumulación: valor medio o mediana, se producen 3 x 2 x 2 x 9 x 2 = 216 características).

Frecuencia fundamental: Para darle la posibilidad al sistema de clasificación de compensar las influencias dependientes de la frecuencia fundamental, se usan el valor medio y mediana de la frecuencia fundamental (logarítmica) (se producen 2 características).

Variabilidad de la sonoridad: Para caracterizar las fracciones de voz chirriantes, se calculan distintas medidas de la variabilidad de la sonoridad (intervalos: 0 a 50 ms, 0 a 100 ms, 0 a 200 ms, fase completa, 200 ms hasta final de frase; sonoridad: solo sonoro o pseudo-período-síncrono; medidas: desviación estándar, errores de las rectas de regresión, índice de variabilidad por parejas, aceleración absoluta media, se producen 5 x 2 x 4 = 40 características;

para la sonoridad independiente de sonoro / sordo -100 a 0 ms, 0 a 50 ms, 0 a 100 ms, 0 a 200 ms, frase completa, 200 ms hasta final de frase; se producen 6 x 4 = 24 características).

Variabilidad sonora / sorda: Igualmente para fracciones de voz chirriantes se caracteriza con qué frecuencia se rompe la voz entretanto, es decir, está presente una señal vocal sorda (fracción de las ventanas de análisis sonora sobre el intervalo -100 a 0 ms, 0 a 50 ms, 0 a 100 ms, 0 a 200 ms, frase completa, 200 ms hasta final de frase; se producen 6 características).

Armonicidad: Para la caracterización de la calidad de la voz se calculan las medidas de la armonicidad (intervalos 0 a 10 ms, 0 a 50 ms, 0 a 100 ms, 0 a 200 ms, frase completa, 200 ms hasta final de frase; solo se cuenta cuando sonoro; armonicidad de correlación cruzada c conforme a la frecuencia fundamental log((c + 0,1) / (1,0001 - c)); se producen 6 características).

Características espectrales 1: Para la caracterización de la calidad de la voz o del sonido hablado se admite la energía en las bandas de frecuencia (24 bandas de Mel, toma de logaritmo; normalizado con valor medio (después de toma de logaritmos) sobre segmentos sonoros); promediado sobre los intervalos 0 a 10 ms, 0 a 50 ms, 0 a 100 ms, 0 a 200 ms, 10 a 50 ms, 10 a 100 ms, 10 a 200 ms; se producen 24 x 7 = 168 características.)

Las bandas de Mel se producen como resultado intermedio en el cálculo de las características de reconocimiento de voz estándar, véase abajo también por ejemplo http://siggigue.github.io/pyfilterbank/melbank.html. Se trata de las energías sumadas (antes de la toma de logaritmo), ponderada con filtros triangulares, que obedecen a una escala orientada.

Características espectrales 2: Para la caracterización adicional de la calidad de la voz o del sonido hablado se usan las características a corto plazo estándares de un sistema de reconocimiento de voz (Mel-Frequency Cepstral Coefficients (MFCC, s. S. B. Davis and P. Mermelstein, Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, nº 4, pp. 357-366, 1980.); 13 MFCC de 24 bandas Mel; normalizada con el valor medio sobre segmentos sonoros; intervalo como arriba; se producen 13 x 7 = 91 características).

Características espectrales 3: Características espectrales 1 +2 en el momento inmediato del uso; se producen 24 + 13 = 37 características.

Sonoridad antes del uso: Para la modelización de una aspiración audible (que se debería evitar según el comportamiento objetivo en la Fluency Shaping) se caracteriza la sonoridad antes del uso (intervalos: 100 ms antes del uso o hasta 200 ms (depende del tiempo disponible antes del uso); normalización como en sonoridad del uso inmediato, se producen 2 x 6 = 12 características).

Espectro antes del uso: Para la misma finalidad se admiten las energías de las bandas de frecuencia de Mel antes del uso (dos intervalos como arriba; normalización: todos los segmentos sonoros, o valor medio de los primeros 50, 100 o 200 ms tras el uso; se producen 2 x 4 x 24 = 192 características).

3. Clasificación

La reproducción del vector de características de 908 dimensiones sobre la clase de objetivo (p. ej. uso suave / duro) se realiza con ayuda de máquinas de vectores de soporte. Este procedimiento impulsado por datos se destaca por la potencia sobresaliente también en el caso de datos de entrada de grandes dimensiones. Para resolver un problemas de más clases con este procedimiento de clasificación se aplica el esquema “uno-contra-resto”. Todas las características se normalizan en primer lugar individualmente a la desviación estándar 1; luego se normaliza la longitud euclidea promedio de un vector de características a 1. De este modo se puede seleccionar de forma muy sencilla el metaparámetro C (constante a C = 0,001). Se usa una función de Kernel lineal. Para la estimación precisa de probabilidades de clase se transforman de forma apropiada las distribuciones de funciones de decisión.

La base de las máquinas de vectores de soporte es la solución de un problema de optimización no lineal, que se destaca en la aplicación por la robustez especial en la clasificación. Los datos de entrada son la anotación y las características del locutor de entrenamiento. Mediante la función de Kernel lineal usada, la fase de clasificación acaba en una función lineal afín sencilla. Se usan bibliotecas de programas estándares para el cálculo.

Los resultados se transforman con la finalidad de la estimación de la probabilidad mediante una función de Sigmoid, véase http://scikit-learn.org/stable/modules/calibration.html. Aquí también se usan bibliotecas de programas estándares para el cálculo.

4. Datos

En el marco de cursos regulares de la KST se han acumulado grabaciones de voz de tartamudos. A este respecto, los locutores deberían hablar distintos ejercicios: vocales sostenidas, sílabas individuales, palabras monosílabas, bisílabas y trisílabas, y finalmente frases cortas y largas. Para la digitalización se usan auriculares económicos.

Existen 3586 grabaciones de 16 locutoras femeninas y 49 masculinos en la entorno de en conjunto 5,0 horas. Todos los oradores son hablantes nativos de alemán.

4.1. Anotación

En el material se han realizado notas por cinco terapeutas, en el sentido de que se han implementado respectivamente adecuadamente distintos aspectos del comportamiento objetivo en el Fluency Shaping. Aquí solo interesan las anotaciones para el uso de la voz, que se ha valorado respectivamente con “bien” (es decir, suave), “dura”, “aspirada”, “chirriante” o “afectada con oclusión glotal”. El material se ha representado en orden aleatorio; por grabación existen en el corte anotaciones de 2,8 terapeutas (es decir, no cada terapeuta ha valorado todas las grabaciones).

4.1.1. Acuerdo entre evaluadores

Para valorar el bien definido de criterios (o la dificultad de la tarea de anotación) se ha considerado cuan adecuadamente concuerdan las anotaciones de los terapeutas individuales. Se ha usado por parejas, es decir, se ha comparado respectivamente un terapeuta con otro. A este respecto, se debe atender a que esto es una estimación pesimista de la calidad, dado que al primer terapeuta se le atribuyen parcialmente los errores del otro como presuntos errores.

La tabla 1 muestra la matriz de confusión para usos de vocales. El reconocimiento de los tipos de errores individuales se puede caracterizar también con ayuda de cuotas de aciertos y falsas alarmas. Estas están reunidas en la tabla 2. En resumen se puede decir que los criterios se representan como bien definido, dado que la concordancia es muy al azar. Por ejemplo la cuota de aciertos “no suave respecto a suave” con 71,5 % con 22,7 de falsa alarma (véase tabla 2) es esencialmente mayor que la concordancia al azar a esperar aquí de 22,7%.

Tabla 1: Matriz de confusión para la anotación de usos de vocales en la comparación por parejas de los terapeutas. Dado que a este respecto cada uno de los 1472 usos se ha contado doble, se producen en conjunto 1638 + 470 + 203 + 565 + 68 = 2944 = 1472 x 2 casos.

Anotador 1	# Casos	Anotador 2 [%]				
		suave	duro	aspirado	crujiente	oclusión glotal
suave	1638	77,3	7,1	6,6	6,9	2,1
duro	470	24,7	37,4	2,6	31,1	4,0
aspirado	203	53,2	5,9	28,6	10,8	1,5

Anotador 1	# Casos	Anotador 2 [%]				
		suave	duro	aspirado	crujiente	oclusión glotal
crujiente	565	20,0	26,0	3,9	49,6	0,5
oclusión glotal	68	51,5	27,9	4,4	4,4	11,8

Tabla 2: Cuota de aciertos / falsa alarma en % para el reconocimiento de distintos tipos de errores durante el uso de la voz (usos vocales; comparación por parejas de los terapeutas).

	no suave	duro	aspirado	crujiente	oclusión glotal
vs. suave	71,5 @ 22,7	37,4 @ 7,1	28,6 @ 6,6	49,6 @ 6,9	11,8 @ 2,1

4.2. Datos planteados

Adicionalmente los terapeutas (seis locutoras femeninas, uno masculino; todos hablantes nativos en alemán) hablaron partes del material todavía en dos variantes, una vez con habla normal, y otra vez con la técnica de voz suave, unida de Fluency Shaping. Aquí están presentes 5338 grabaciones en el entorno de en conjunto 4,7 horas.

Bajo la suposición simplificadora que se han generado de este modo usos de la voz duros y suaves correspondientes, se ha prescindido de realizar anotaciones en esta parte del material. Esto se ha confirmado por una inspección de tipo muestra del material al menos para los usos de la voz con vocales. En contraposición a los datos arriba mencionados de tartamudos, donde están contenidos muchos casos límite, en este material planteado está presente respectivamente un marcado claro, prototípico de la propiedad de uso de voz “suave” o “duro”.

5. Experimentos y resultados

La fiabilidad a esperar del sistema automático para la clasificación de usos de la voz se ha estimado experimentalmente. Para ello se entrena respectivamente con una parte de los hablantes y se testea con los hablantes restantes (es decir, los parámetros del sistema de clasificación se estiman con ayuda de hablantes anotados en el conjunto de entrenamiento, y la exactitud se mide con ayuda de hablantes anotados en el conjunto de test). En conjunto se obtiene un valor estimado conservador de la exactitud del sistema definitivo, que se entrena con ayuda de todos los hablantes. Se usa una validación cruzada doble; los resultados se promedian sobre 10 pasos con distintas clasificaciones de hablantes de entrenamiento y test. En el entrenamiento se usa una validación cruzada interna (igualmente independiente del hablante, igualmente doble) para la calibración de las probabilidades. Si están presentes varias anotaciones para un uso, se usan todas compitiendo (tanto en el entrenamiento como también test). De este modo la exactitud producida es directamente comparable con la concordancia entre evaluadores de la sección 4.1.1 e igualmente pesimista (ya que también al sistema se le atribuyen parcialmente los errores de los terapeutas como presuntos errores).

Se evalúan solo usos de vocales. Solo se evalúan grabaciones en las que el número de las frases segmentadas automáticamente (véase sección 1) concuerda con la anotación de los terapeutas. (Motivo: por motivos de costes se prescinde de una segmentación manual del material; mediante el modo de proceder descrito es posible una asociación sencilla, aproximada de las frases segmentadas de forma automática y manual por el índice de la frase respectiva.) El sistema de clasificación emite probabilidades para los distintos usos de la voz. Según desde que probabilidad se anuncia un error de pronunciación, se produce un sistema más estricto (con cuota de acierto más elevada pero también cuota de falsa alarma más elevada) o un sistema más seguro (cuota de acierto más baja pero también menos falsa alarma). Esto se puede representar en una así denominada curva característica operativa del receptor (curva ROC).

5.1. Insuficiencia de sonoridad

Para verificar la hipótesis de la insuficiencia del uso solo de la sonoridad como criterio de clasificación para la valoración de usos de la voz, en particular para su clasificación como suave, se ha examinado la fiabilidad del sistema de clasificación en el uso de distintos subconjuntos de las características de la sección 2.2:

Desarrollo de sonoridad: amplitud, amplitud local, sonoridad del uso, sonoridad del uso directo y subida de la sonoridad (102 características); frecuencia fundamental / múltiplo armónico (216 características);

Resto: Ruido de fondo, bajada de sonoridad, frecuencia fundamental, variabilidad de la sonoridad, variabilidad sonora / sorda, armonicidad, características espectrales 1-3, sonoridad antes del uso, y espectro antes del uso (590 características).

Los resultados se encuentran en la fig. 3 en las representaciones a a c. Para las características solo basadas en el desarrollo de sonoridad no se produce un buen poder de reconocimiento: Por ejemplo, el área bajo la curva (Area under the Curve, AUC) para “suave” versus “no suave” con 0,424 es incluso menor que el valor de la línea base al azar de 0,5.

Los experimentos, con las grabaciones prototípicas de las terapeutas (sección 4.2), muestran que esto no depende de por ejemplo un cálculo diferente de los parámetros de sonoridad, sino de la dificultad de la tarea: cuando el sistema se testea respecto a estos, se producen ya resultados muy buenos con la sonoridad. Por ejemplo se produce para “suave” versus “no suave” un AUC de 0,906 (con todas las características se produce el número todavía ampliamente mejorado de 0,973).

Solo cuando se usan las características adicionales se producen también buenos resultados en los datos realistas de tartamudos. En este caso se debe mencionar (véase la fig. 3b) que los análisis que se basan en la frecuencia fundamental o múltiplos armónicos, a usar en cualquier caso según la invención, son especialmente apropiados para el reconocimiento de usos chirriantes, y los restantes (véase la fig. 3c) para el reconocimiento de usos duros. Pero también se puede reconocer que ya una reducción de los parámetros observados sobre los valores de energía sacados según la invención de la frecuencia fundamental y múltiplos armónicos durante el uso de la voz produce una mejora, significativa respecto a la simple observación de la sonoridad, del comportamiento de reconocimiento de un procedimiento realizado con estos parámetros y especificaciones. Las ventajas de los grupos de parámetros respectivos se combinan en el sistema con todas las características (véase la fig. 4).

Así se pudo mostrar que el sistema de reconocimiento de patrones propuesto con la invención puede obtener un reconocimiento fiable de usos de la voz suaves / no suaves, que se puede usar p. ej. para la terapia de tartamudeo según el Fluency Shaping. Esto también es aplicable para la aplicación en condiciones realistas. Se ha mostrado que los enfoques hasta ahora (como por ejemplo según el documento US 4,020,567) no proporcionan esto (sino que solo lo hacen para datos prototípicos, claramente marcados que no son realistas para la aplicación). La fiabilidad del sistema está en el mismo orden de magnitud que la de un terapeuta. Por ejemplo, para la diferenciación de suave vs. no suave con 22,7% de falsa alarma se produce una cuota de acierto del 58,5% para el sistema, mientras que un terapeuta de media logra el 71,5%.

REIVINDICACIONES

1. Procedimiento para la valoración de una calidad de un uso de la voz de un hablante, donde se detecta una señal vocal acústica del hablante y se convierte en una señal vocal digital, donde la señal vocal digital se analiza en su desarrollo temporal, para

a. determinar en el desarrollo temporal de la señal vocal digital un momento del uso de la voz del hablante,
b. determinar una frecuencia fundamental de la señal vocal en el momento del uso de la voz,
c. determinar a partir de la señal vocal digital en un intervalo de tiempo predeterminado desde el momento del uso de la voz de forma resuelta temporalmente el desarrollo de la energía contenida en la señal vocal a la frecuencia fundamental;
d. determinar a partir de la señal vocal digital en el intervalo de tiempo predeterminado de forma resuelta temporalmente el desarrollo de la energía contenida en la señal vocal en al menos un múltiplo armónico de la frecuencia fundamental;
e. determinar el desarrollo temporal de la relación de las energías determinadas en las etapas c y d,

caracterizado porque se concluye un uso de la voz suave cuando, en el intervalo de tiempo, la relación de las energías determinada según la etapa e. anterior se domina en primer lugar por la energía en la frecuencia fundamental y solo en el desarrollo posterior del intervalo de tiempo predeterminado en un lapso de tiempo Δt se desplaza la relación de las energías en beneficio de la(s) energía(s) en el/los múltiplo(s) armónico(s) de la frecuencia fundamental.

2. Procedimiento según la reivindicación 1, **caracterizado porque** para la determinación del instante del uso de la voz en la etapa a. se usa un algoritmo RAPT según David Talkin.

3. Procedimiento según cualquiera de las reivindicaciones anteriores, **caracterizado porque** el intervalo de tiempo predeterminado presenta una longitud de 7,5 a 200 milisegundos, en particular de 50 a 100 milisegundos.

4. Procedimiento según cualquiera de las reivindicaciones anteriores, **caracterizado porque** el lapso de tiempo Δt está entre 50 y 100 milisegundos y este se evalúa como criterio para un uso de la voz suave.

5. Procedimiento según cualquiera de las reivindicaciones anteriores, **caracterizado porque** junto a un análisis de la relación de las energías determinada en la etapa e. se recurre a otros parámetros derivados de la señal vocal digital para determinar si se trata de un uso de la voz suave.

6. Programa de tratamiento de datos para la valoración automatizada, asistida por ordenador de una calidad de un uso de la voz de un hablante, con

a. un módulo de análisis del uso de la voz, que está establecido para determinar a partir de un desarrollo temporal de una señal vocal digital obtenida a partir de una señal vocal acústica del hablante un momento del uso de la voz del hablante,
b. un módulo de determinación de la frecuencia fundamental, que está establecido para determinar una frecuencia fundamental de la señal vocal en el momento del uso de la voz,
c. un módulo de determinación de la energía de la frecuencia fundamental, que está establecido para determinar a partir de la señal vocal digital en un intervalo de tiempo predeterminado desde el momento del uso de la voz de forma resuelta temporalmente el desarrollo de la energía contenida en la señal vocal a la frecuencia fundamental;
d. módulo de determinación de la energía de los armónicos que está establecido para determinar a partir de la señal vocal digital en el intervalo de tiempo predeterminado de forma resuelta temporalmente el desarrollo de la energía contenida en la señal vocal en múltiplos armónicos de la frecuencia fundamental;
e. un módulo de determinación de la relación, que está establecido para determinar el desarrollo temporal de la relación de las energías determinadas por el módulo de determinación de la energía de la frecuencia fundamental y el módulo de determinación de la energía de los armónicos y concluir un uso de la voz suave cuando, en el intervalo de tiempo, la relación de las energía se domina en primer lugar por la energía en la frecuencia fundamental y solo durante el desarrollo posterior del intervalo de tiempo predeterminado se desplaza la relación de las energías en beneficio de las energías en los múltiplos armónicos de la frecuencia fundamental.

7. Programa de tratamiento de datos según la reivindicación 6, **caracterizado porque** contiene además un módulo de digitalización para la generación de la señal vocal digital a partir de la señal vocal acústica.

8. Programa de tratamiento de datos según cualquiera de las reivindicaciones 6 o 7, **caracterizado porque** el módulo de análisis del uso de la voz para la determinación del momento del uso de la voz usa un algoritmo RAPT según David Talkin.

9. Programa de tratamiento de datos según cualquiera de las reivindicaciones 6 a 8, **caracterizado porque** está configurado como software de aplicación para un terminal móvil, en particular un teléfono inteligente o una tablet-PC.

10. Equipo de tratamiento de datos asistido por ordenador, **caracterizado porque** contiene un programa de tratamiento de datos según cualquiera de las reivindicaciones 6 a 9.

11. Equipo de tratamiento de datos asistido por ordenador según la reivindicación 10, **caracterizado porque** presenta una entrada para la grabación de una señal vocal acústica.

12. Equipo de tratamiento de datos asistido por ordenador según cualquiera de las reivindicaciones 10 u 11, **caracterizado porque** está configurado como terminal móvil, en particular como teléfono inteligente o tablet-PC.

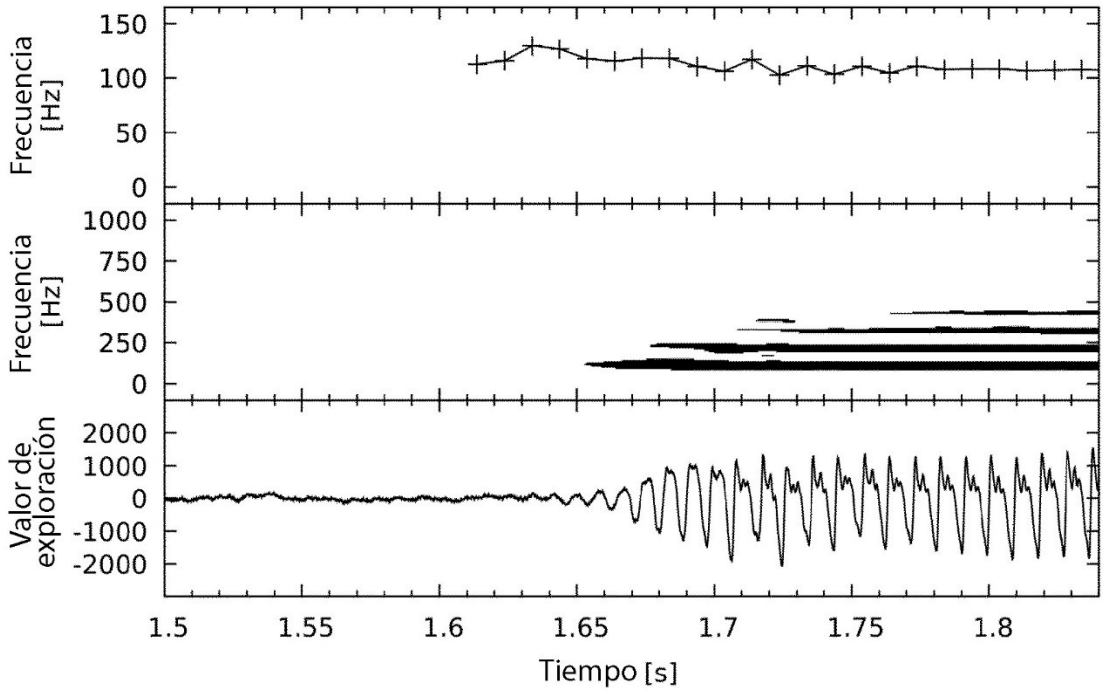
Frecuencia [Hz]
150
100
50
0
Frecuencia [Hz]
1000
750
500
250
0
Valor de exploración
2000
1000
0
-1000
-2000
1.5
1.55
1.6
1.65
1.7
1.75
1.8
Tiempo [s]

Fig. 1

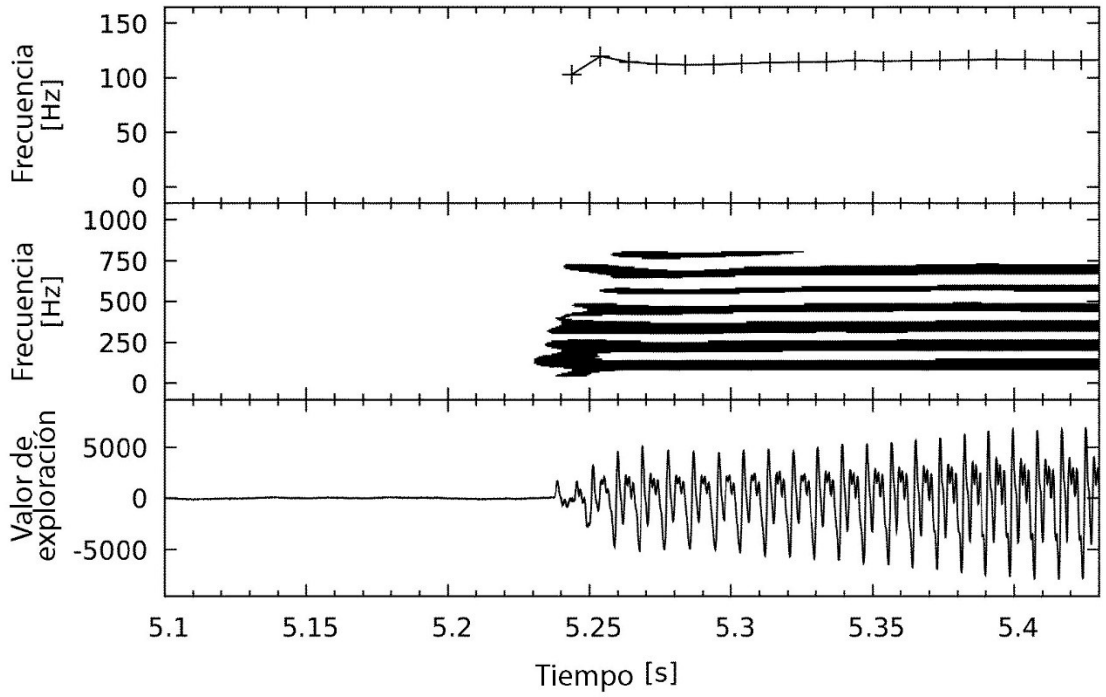
Frecuencia [Hz]
150
100
50
0
Frecuencia [Hz]
1000
750
500
250
0
Valor de exploración
5000
0
-5000
5.1
5.15
5.2
5.25
5.3
5.35
5.4
Tiempo [s]

Fig. 2

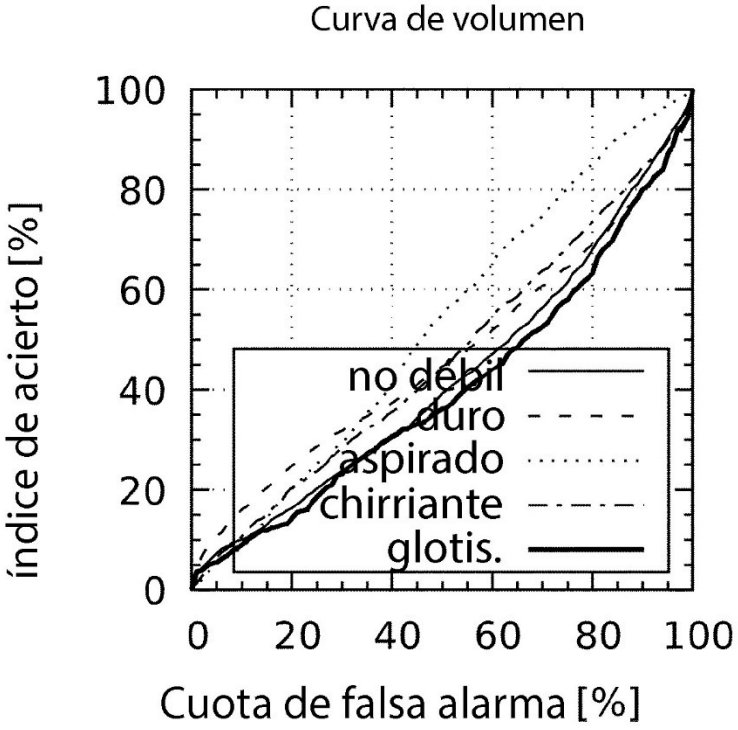
Curva de volumen
índice de acierto [%]
Cuota de falsa alarma [%]
no débil
duro
aspirado
chirriante
glotis.
0
20
40
60
80
100

Fig. 3a

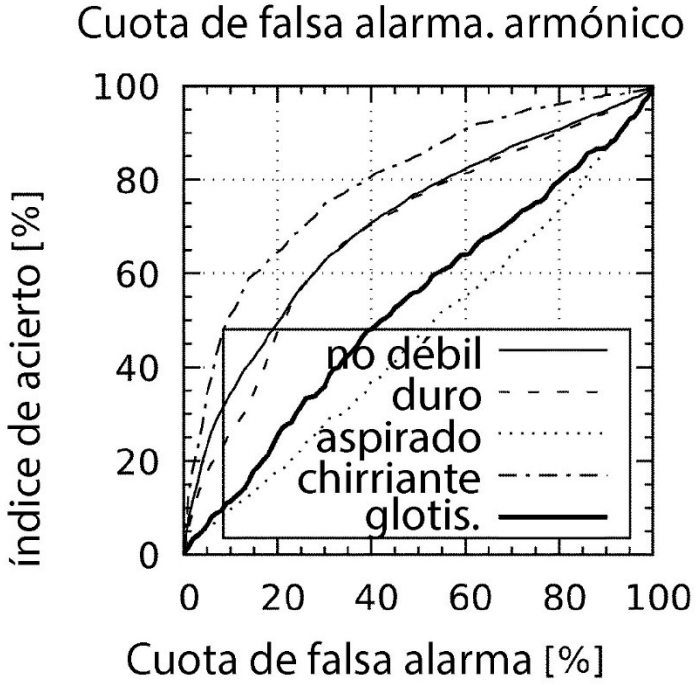
Cuota de falsa alarma. armónico
índice de acierto [%]
Cuota de falsa alarma [%]
no débil
duro
aspirado
chirriante
glotis.
0
20
40
60
80
100

Fig. 3b

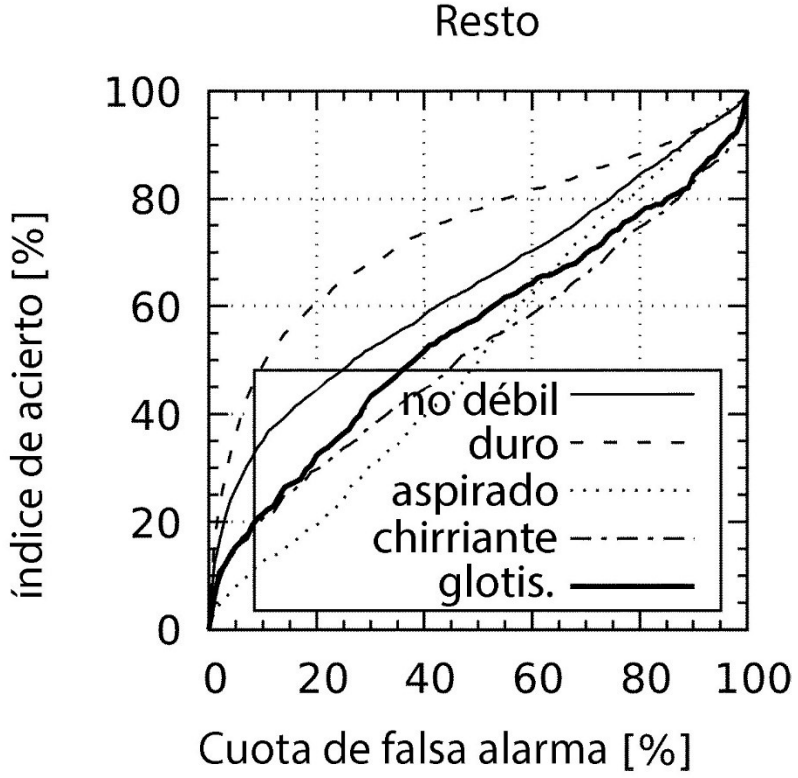
Resto
100
80
60
40
20
0
índice de acierto [%]
0
20
40
60
80
100
Cuota de falsa alarma [%]
no débil
duro
aspirado
chirriante
glotis.

Fig. 3c

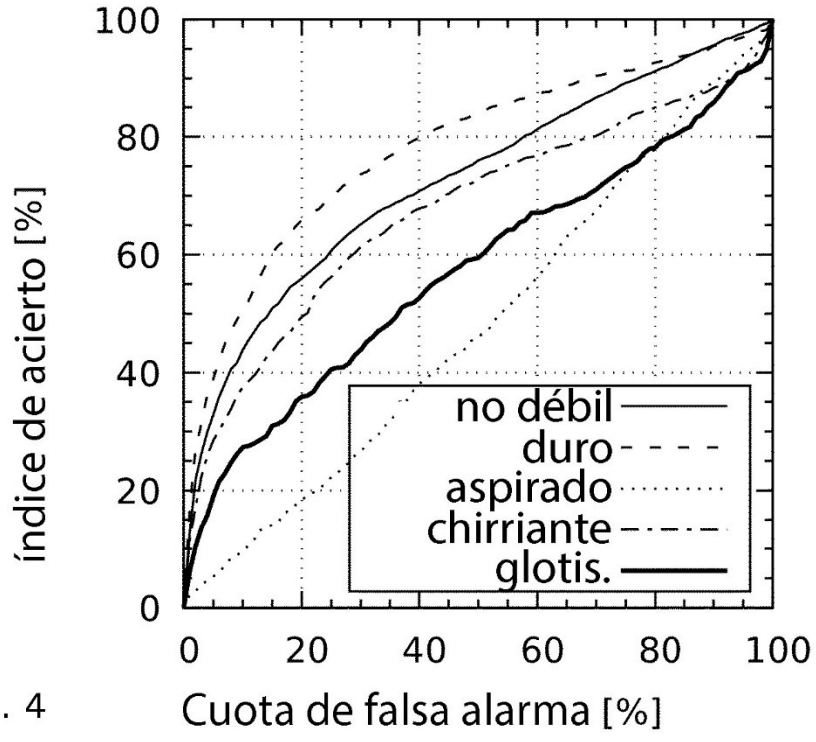
100
80
60
40
20
0
índice de acierto [%]
0
20
40
60
80
100
Cuota de falsa alarma [%]
no débil
duro
aspirado
chirriante
glotis.

Fig. 4