VI Congreso Iberoamericano de Acústica - FIA 2008
Buenos Aires, 5, 6 y 7 de noviembre de 2008

FIA2008-A191

Obtención del Índice STI a partir de la voz

Javier Camusso[(a)],
Cristian Lértora[(b)],
Director: Ing. Federico Miyara[(c)].

(a) Escuela de Ingeniería Electrónica, Universidad Nacional de Rosario.Riobamba 245 bis (S2000EKE) Rosario, Santa Fe. Argentina. E-mail: camussoj@express.com.ar.
(b) Escuela de Ingeniería Electrónica, Universidad Nacional de Rosario.Riobamba 245 bis (S2000EKE) Rosario, Santa Fe. Argentina. E-mail: clertora@arnet.com.ar.
(c) Escuela de Ingeniería Electrónica, Universidad Nacional de Rosario.Riobamba 245 bis (S2000EKE) Rosario, Santa Fe. Argentina. E-mail: fmiyara@fceia.unr.edu.ar.

Abstract

This paper raises the possibility of measuring the speech transmission index (STI) using voice signals instead of the test signals used in traditional methods. The proposal is a new way to determine the speech intelligibility, which uses an utterance that is collected simultaneously at the point of the speaker and at the point of listening. The measurement sought is obtained by working mathematically with these signals. The work focuses on the validation and iterative adjustment of the proposed model. To achieve this goal, an environment simulation in Matlab is developed, allowing the addition of noise and reverberation to audio samples, generating virtual measurements of intelligibility through RASTI model and through the proposed model.

Resumen

Se plantea la posibilidad de medir el índice de transmisión de la voz (STI) utilizando señales de voz en lugar de las señales de prueba utilizadas en los métodos tradicionales. Para ello se propone una nueva modalidad para la determinación de la inteligibilidad de la palabra, la cual utiliza una locución que se recolecta simultáneamente en el punto del locutor y en el punto del escucha. Trabajando matemáticamente estas señales se obtiene la medición buscada. El trabajo se centra en la validación y ajuste en forma iterativa del modelo propuesto. Con este objetivo se crea un entorno de simulación en MatLab, el cual permite adicionar ruido y reverberación a muestras de audio, generando mediciones virtuales de la inteligibilidad a través del modelo RASTI y del modelo propuesto.

1 Introducción

Steeneken y Houtgast presentaron en 1980 un método físico de medida de la inteligibilidad [1], que representa una extensión del método AI (articulation index), basado en la Función de Transferencia de la modulación (MTF: Modulation Transfer Function). El índice resultante se llama índice de Transmisión de la Palabra (STI: Speech Transmission lndex) y se determina a partir de la evaluación de la MTF para 98 combinaciones de bandas de frecuencia y envolventes. A partir del STI, y para reducir el tiempo de proceso, la empresa danesa Brüel y Kjaer patentó el RASTI (Rapid Speech Transmission lndex), que analiza 9 frecuencias de modulación, 4 para la banda de 500 Hz y 5 para la de 2 kHz.

Estos métodos adolecen de un problema: Sólo pueden determinar una calificación bajo condiciones controladas y estables del recinto, y sus conclusiones sólo son válidas bajo dichas condiciones. Esto no permite la evaluación en tiempo real de la inteligibilidad de la palabra frente a variaciones de las condiciones del recinto (ruido, apertura o cierre de ventanas, etc.).

Se plantea aquí la posibilidad de medir la inteligibilidad utilizando señales de voz, superando así las falencias de los métodos mencionados, es decir, permitiendo la medición de la inteligibilidad en condiciones totalmente realistas, por ejemplo en vivo (condiciones no controladas).

2 Métodos tradicionales basados en la MTF

Se exponen a continuación someramente los métodos STI y RASTI cuya filosofía toma por base este trabajo.

2.1 STI

El método STI es un método objetivo de medición de la inteligibilidad con un rango de valores comprendidos entre 0 (inteligibilidad nula) y 1 (inteligibilidad óptima).

Se basa en las contribuciones (apropiadamente ponderadas) de la relación señal a ruido efectiva (SNR) en un grupo de 14 frecuencias de modulación en 7 bandas de octava (98 puntos) distribuidas en el rango que ocupa la voz.

Para obtener la relación señal a ruido efectiva se emplean una emisión y una recepción especiales. El emisor (Voz artificial) envía una **señal de prueba** con información en las frecuencias típicas de la voz y fluctuaciones de intensidad dentro de la misma, de forma que simula un locutor. El receptor se coloca en la posición del oyente, y mide como se ha modificado la señal emitida. La alteración de la señal (la reducción de la modulación) en la posición del oyente se cuantifica a través de la función de transferencia de modulación (MTF) para las 98 frecuencias diferentes, se pondera apropiadamente y se expresa en función de una relación señal a ruido aparente, la cual es acotada y convertida en el valor STI.

2.2 RASTI

Para la mayoría de las situaciones de auditorios actuales, el sistema de 98 puntos de medida constituye un número innecesario de puntos de análisis. De este modo, para una evaluación más rápida de las condiciones de auditorios Brüel y Kjaer definió un procedimiento de medida más veloz, basado en un subconjunto de los 98 puntos de STI. El método se llama RASTI y se basa, al igual que el STI, en la obtención de la MTF tras la utilización de una emisión especial, una señal de prueba basada en un número menor de frecuencias de modulación. El análisis se restringe sólo a dos bandas de octava de frecuencias centrales 500 Hz y 2 kHz, con cuatro (1 Hz, 2 Hz, 4 Hz y 8 Hz) y cinco (0,7 Hz, 1,4 Hz, 2,8 Hz, 5,6 Hz y 11,2 Hz) frecuencias de modulación respectivamente.

3 Presentación del Modelo Propuesto

La idea que se plantea apunta al desarrollo de un método de medición que utilice las señales de audio provenientes de dos micrófonos, uno ubicado frente al locutor, y el otro en el punto en el que se está midiendo. Procesando estas señales, se obtiene la medición de inteligibilidad.

Figura 1. Esquema de la obtención de señales.
x = Señal "Limpia",
y = Señal "Sucia".

Este método constaría de varias etapas:

1- Alineación temporal. Deben alinearse temporalmente las señales x e y antes de ser procesadas. Esto puede realizarse con algún algoritmo automático, o utilizando el dato de la distancia entre micrófonos.

2- Recorte de silencios. Se deben quitar las porciones silenciosas de la muestra ya que no poseen información útil y podrían desvirtuar la medición. Para esto se analiza la señal x y se genera un patrón de recorte que se aplica a ambas señales.

3- De manera análoga al método RASTI, se miden las componentes espectrales de las señales en determinadas frecuencias y se determina la reducción de modulación. Con esto se obtiene un primer valor del índice STI.

4- Se corrige el valor anterior con una función de adecuación. La misma deriva de la comparación de mediciones realizadas con este método y con RASTI en un universo de muestras suficientemente amplio.

En lo sucesivo, se denominará VRASTI (Voice RASTI) al método descrito. El Presente trabajo pretende brindar una primera aproximación al método propuesto y analizar su eficacia con el fin de evaluar la factibilidad de la propuesta. Para lograr este objetivo, se desarrolló un entorno de simulación que permite emular la acción de distintos recintos sobre señales de voz, obteniendo así el universo de muestras necesario para la evaluación.

Para las simulaciones e implementación de los algoritmos de cálculo de los distintos métodos, se eligió como plataforma de desarrollo al programa MatLab.

3.1 Implementación de RASTI

Como primer paso se genera la **señal de prueba** siguiendo explícitamente las instrucciones del estándar IEC [2]. En el diagrama de bloques de la figura 2 se muestra cómo se da forma a la modulación en intensidad de la envolvente, luego se convierte a presión sonora vía la aplicación de una raíz cuadrada y finalmente se aplica al ruido-portadora apropiadamente filtrado en las dos bandas de octava.

Luego, se implementa el método RASTI como se muestra en la figura 3.

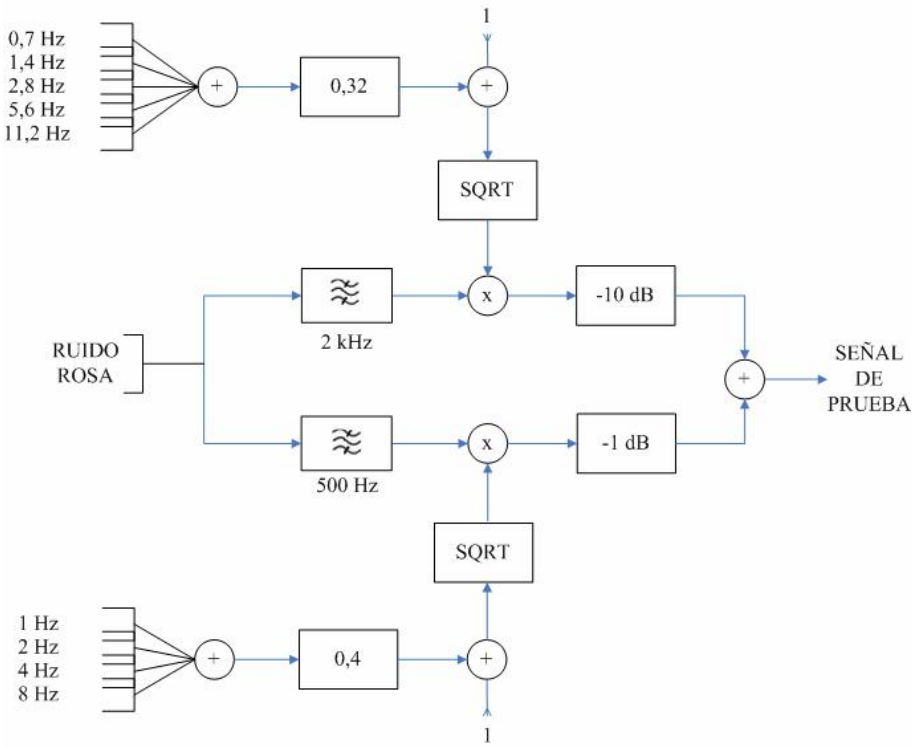

Figura 2. Diagrama de bloques de la generación de la señal de prueba para el método RASTI.

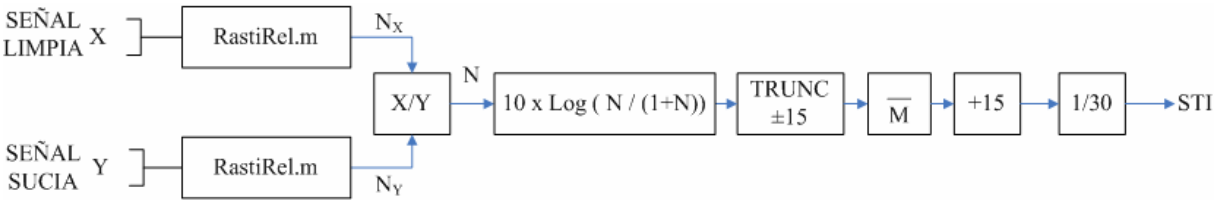

Figura 3. Implementación de método RASTI.

El primer procesamiento de las señales lo realiza el bloque **"RastiRel"**, cuyo detalle se muestra en figura 4. En este bloque se obtiene el espectro de baja frecuencia propio de la intensidad (se pasa de presión sonora a intensidad elevando al cuadrado) por medio de una FFT. Se toma su valor para cada una de las frecuencias de modulación y se lo relaciona con la intensidad en cero (proporcional a la intensidad de envolvente para cada frecuencia). Esto se realiza para las bandas de las octavas centradas en 2 kHz y 500 Hz obteniéndose 9 valores de índice de modulación (N).

Estos índices se obtienen para las señales de entrada (N_X) y salida (N_Y) al medio en evaluación. Realizando el cociente índice a índice se obtienen los 9 valores N (figura 3) propios de la función de transferencia de modulación (MTF). Se aplica una transferencia alineal para convertir los valores a una escala en dB, se los acota al rango [-15 +15] y se los unifica aplicando la media. Finalmente se ajusta a un rango final [0 1] obteniéndose así el valor STI.

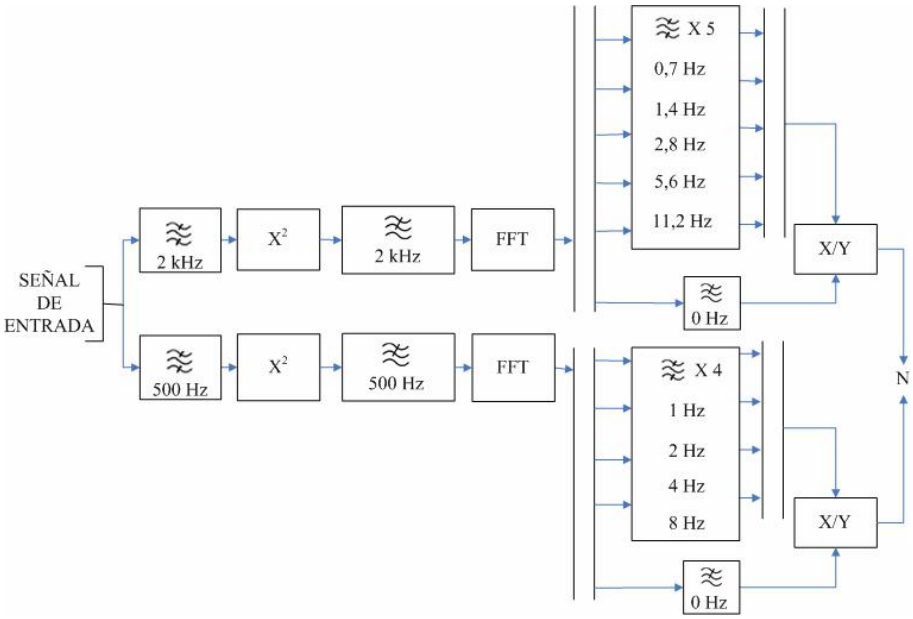

Figura 4. Etapa de obtención de reducción en los índices de modulación m mediante el método RASTI. Implementado con el nombre RastiRel.

3.2 Implementación de VRASTI

Se modeló el VRASTI con una estructura análoga al RASTI cambiando el modo en que se extrae la información para medir la pérdida de modulación y se antepuso un algoritmo que elimina los silencios prolongados entre palabras dentro de la cadena de habla.

La necesidad de cambiar la manera de medir las energías con que se calcula la pérdida de modulación la suscita el hecho de que las frecuencias de modulación de la voz no están concentradas en determinados puntos del espectro como sí sucede en la señal de prueba de RASTI, por lo que se considera en VRASTI una banda y no un único punto de frecuencia.

La eliminación de silencios es necesaria porque el presente trabajo está enfocado a lograr un procesamiento en tiempo real y no es posible conocer a priori si el discurso a ser analizado estará o no libre de silencios. La ausencia de palabra deja a éste modelo sin su "señal de prueba", por lo que es imprescindible la eliminación de estos silencios.

3.2.1 Extracción de Silencios

Para este trabajo se utilizó un algoritmo de detección de actividad de habla recursiva (RVAD) que estima de forma automática, mediante un procedimiento recursivo, los períodos de tiempo en los que no está presente la voz. A éste se le realizó un suavizado para evitar distorsiones mediante la aplicación del método de "**recursión de polo único bilateral**" [3], que utiliza un doble valor para λ, uno de ataque y uno de caída.

Se toman la señal limpia y su versión alterada por el ambiente bajo análisis; de la primera con el RVAD se obtienen los espectros de señal (SS) y de ruido (NN), correspondientes trama a trama a la señal origen S_0 y se define un umbral. Toda trama de S_0 para la que la diferencia SS-NN no supera al umbral mencionado se considera silencio. Además, se define un tiempo mínimo para la eliminación de tramas. Sólo los silencios detectados que superen en duración a este mínimo, serán considerados silencio. Finalmente se agrega una curva de ataque y caída para no introducir recortes abruptos y se ha obtenido un patrón de presencia de silencios.

En base a este patrón se procesan tanto la versión limpia como la alterada devolviéndose dos señales procesadas más cortas, sincrónicas entre sí, y sin presencia de tramas silenciosas.

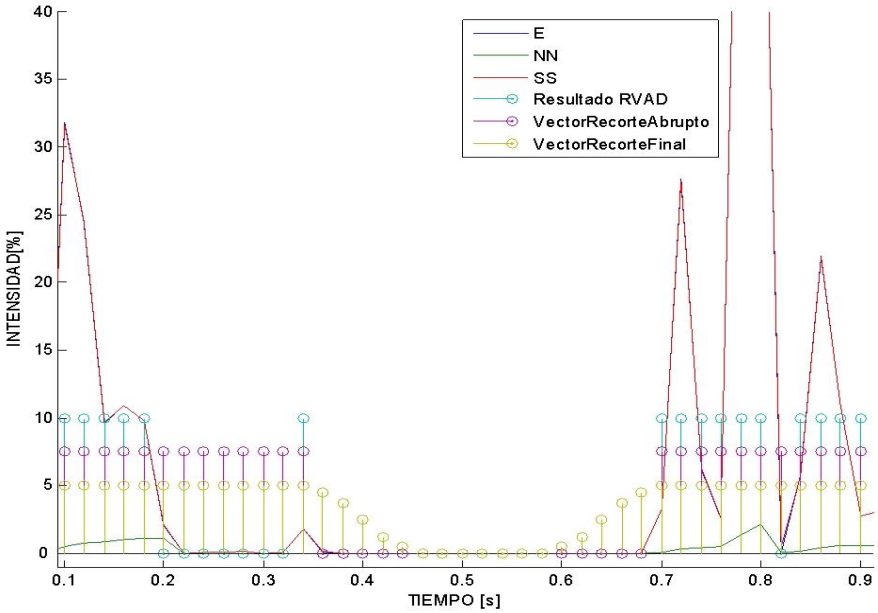

Figura 7. Comportamiento del algoritmo encargado de recortar silencios.

3.2.2 Obtención del índice STI con VRASTI

Este modelo toma la estructura del bloque **"RastiRel"** y lo adapta en función al hecho de que la señal de entrada ya no es una simulación, sino una locución real. Se diseñó una serie de ventanas espectrales de octava FFT con transferencia tipo ventana Hanning para aplicar alrededor de cada frecuencia moduladora. El valor de la integral del contenido de cada ventana, relativo a la componente de continua en la banda a la que pertenece la moduladora en cuestión se utiliza en el cálculo de la pérdida del índice de modulación. El bloque se denominó **"VRastiRel"**.

El resto de la estructura es análoga a la versión RASTI. En la figura 5 se muestra el procesamiento completo que culmina en la obtención del valor STI preliminar. Este valor no será el definitivo ya que luego será ajustado por la función de adecuación. Cabe aclarar que las señales "limpia" y "sucia" (figura 4) han sufrido previamente la extracción de silencios.

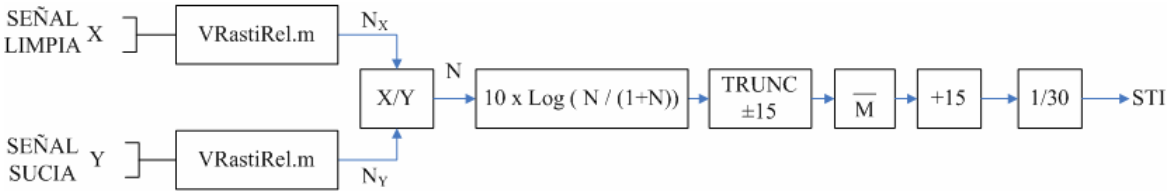

Figura 5. Implementación de modelo VRASTI.

4 Planteo de Simulaciones

Se crearon herramientas capaces de variar el ruido rosa y la reverberación aplicada a voces masculinas y femeninas de 4 locutores con el objeto de gobernar estos efectos y abarcar gran parte del rango de variación de STI, variando los parámetros "Tiempo de Reverberación" (TR) y "Relación Señal a Ruido" (SNR). El objetivo es exponer al modelo VRASTI a este ensayo e ir comparando sus resultados contra los arrojados por RASTI y en base a esta comparación generar un polinomio de adecuación que aporte un ajuste final a los valores arrojados por VRASTI. A continuación se detallan las herramientas mencionadas.

4.1 Duración de las Muestras

Se realizaron mediciones con VRASTI sobre muestras de voz de diferente longitud, determinándose que las mediciones se estabilizan para duraciones mayores a 28 s [7]. Por lo anterior se decidió trabajar con señales de alrededor de 60 s de duración, extraer sus silencios y recortarlas a 30 s.

4.2 Implementación de Ruido Rosa

Se creó una función que adiciona ruido rosa. A la misma se la alimenta con una señal y una relación señal a ruido (SNR) en dB, y retorna la señal con ruido rosa adicionado, con la SNR deseada.

4.3 Implementación de Reverberación

Para aplicar reverberación a un discurso grabado se plantea una respuesta al impulso para una sala ficticia y luego se convoluciona con la señal de voz. Se varía la forma de la respuesta al impulso una y otra vez aplicándolo a la cadena de audio con el objetivo de obtener un cambio proporcional del valor STI correspondiente a cada caso.

Las reflexiones tempranas se modelaron con cuatro impulsos de intensidades variadas, acotados por un factor dado en porcentaje (del impulso unitario). La separación entre impulsos es variada, función de la forma de la sala y varían proporcionalmente al tiempo de reflexiones tempranas (Trt)[1].

Las reflexiones tardías o cola reverberante se modelaron con un ruido blanco de amplitud máxima o inicial proporcional al impulso unitario (porcentual) modulado por una exponencial decreciente.

Se llamó tiempo de reverberación (TR) a la suma de los tiempos de cola (Tcola) y de reflexiones tempranas (Trt).

En particular para los ensayos realizados durante este trabajo, se fijaron la mayoría de parámetros y proporciones mencionadas y solo se varió el largo de la cola (se fijó Trt en 60 ms y se varió TR), lo que equivale a cambiar la absorción de las superficies de la sala pero no su forma.

Procediendo de esta manera se obtienen buenos resultados, tanto en la variación de STI como al escuchar los discursos afectados por el modelo.

5 Función de adecuación

Se eligió como función de adecuación un polinomio de orden 3, que utiliza sólo los datos del conjunto de muestras en que el valor RASTI es mayor a 0,25, ya que el ajuste por debajo de este valor carece de importancia para los resultados.

Dado que se utilizan para las simulaciones distintos locutores, se plantean dos alternativas de utilización del modelo VRASTI.

La primera es ajustar los resultados obtenidos para un locutor en particular. En este caso se toma en forma aleatoria la mitad de los pares de medición VRASTI-RASTI y se obtiene el polinomio que minimiza el error (diferencia entre mediciones). Esto plantea un entrenamiento del método para este locutor en particular[2]. La otra mitad de las muestras se utiliza para la evaluación del ajuste realizado.

---

[1] En el modelado de las reflexiones tempranas no se puso énfasis en que representasen la geometría de alguna sala en particular porque para el fin que perseguimos solo necesitamos variar sus características acústicas y nos alcanza con introducir cambios solo en la cola.

[2] Un dispositivo que implemente esta idea debería ofrecer la posibilidad de ser entrenado por el locutor que se utilizará como fuente en la medición. Este dispositivo debería procesar las simulaciones aquí mencionadas para obtener el polinomio de adecuación.

La otra alternativa evalúa el comportamiento del modelo para un locutor genérico. En este caso se toman las mediciones de tres locutores para entrenamiento del polinomio y las del cuarto para su evaluación.

En la exposición de resultados las dos alternativas mencionadas se denominaran como "**Locutor Particular**" y "**Locutor General**" respectivamente.

6 Modificaciones al Modelo Propuesto

Durante el desarrollo de este trabajo se observó en el dominio de la frecuencia que al adicionar reverberación[3] a las tramas de audio, dado el aporte de bandas vecinas, además de la esperada reducción de modulación se tiene un incremento en la energía en las cercanías de la frecuencia moduladora en estudio. Esto se puede observar en la figura 6 donde para una mejor apreciación del fenómeno se usó la señal de prueba RASTI en lugar de la voz.

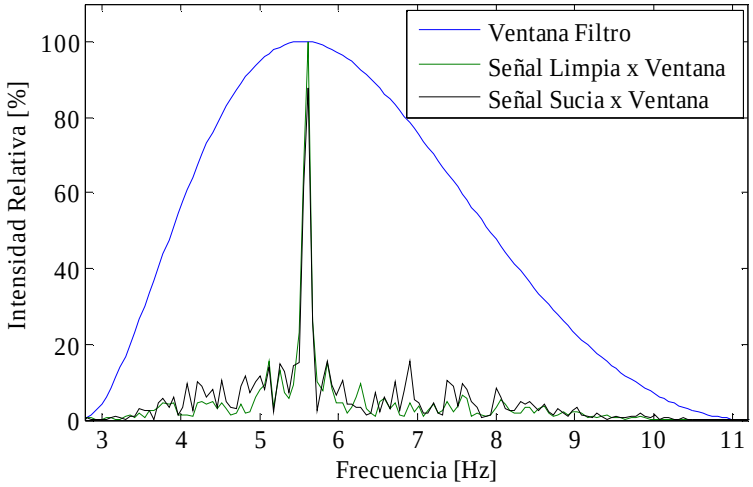

Figura 6. Espectro filtrado en las vecindades de 5,6 Hz. La señal sucia fue alterada por una reverberación con TR = 100 ms. La relación entre las energías de las dos señales supera a la unidad.

Dado que el modelo que se plantea toma una banda del espectro y no un único punto, este fenómeno afecta las mediciones, provocando una sobre valoración de los índices de modulación, con lo que se obtienen relaciones que superan a la unidad. Como el sistema propuesto considera que esto no ocurrirá pueden resultar truncados valores claves.

Buscando paliar este inconveniente se propusieron distintas soluciones que permitan tener en cuenta para la obtención del valor STI estos casos. Se intentó disminuir este efecto de sobre valuación aumentando la selectividad de las ventanas espectrales implementando ventanas Hanning de tercios de octava lo que resultó en una mejora en el comportamiento general del método. Luego se compararon distintos cambios en el procesamiento de los puntos de la MTF (N en figura 4) para obtener el valor STI, como imponer a cada elemento de N un factor de ajuste o eliminar las alinealidades del procesamiento matemático dejando la totalidad del trabajo al polinomio de ajuste. De la comparación entre estas alternativas resultó de mejor desempeño la última con lo que se unieron las mejoras sobre distintas etapas del modelo (ventanas espectrales más selectivas y transferencia lineal) en una nueva versión de VRASTI.

7 Resultados

Se evaluaron diversas modificaciones en el modelo VRASTI resultando entre ellas, tanto para locutores particulares como para locutores generales, de mejor desempeño la implementación de ventanas espectrales de tercio de octava y la conversión de los valores de

[3] También se observó que esto no ocurre al adicionar ruido.

MTF a un único valor STI con la sola aplicación de la media y el polinomio de adecuación. De la unión de estas dos modificaciones surge una **versión mejorada de VRASTI** la que se comparó con la propuesta original para confirmar una mejora en el método. A continuación se exponen y enuncian algunos resultados.

Tanto para el análisis de resultados como para la adecuación polinomial se utilizar las muestras correspondientes a un valor RASTI mayor a 0,25, habiéndose corroborado que el algoritmo VRASTI mantenía el mismo comportamiento por debajo de este valor.

Se evaluaron un total de 1612 ambientes simulados (26 reverberaciones con TR 70 ms hasta 800 ms de a pasos de 30 ms y 62 niveles de ruido rosa con SNR de -30 dB hasta 30 dB de a pasos de 1 dB) sobre discursos de 4 locutores.

En las figuras 7 y 8 se muestran los resultados obtenidos para uno de los casos analizados. Las gráficas de la **izquierda** representan los valores STI obtenidos con RASTI, VRASTI y VRASTI corregido por la función de adecuación. Las de la **derecha** representan la correspondiente discretización de estos valores a la escala cualitativa [4]. En ambos casos los datos se ordenaron en el orden creciente del valor RASTI.

En figura 7 se aprecia la presencia de truncamiento de los valores de VRASTI que sin corrección superaron a la unidad. En la figura 8, para la **versión mejorada de VRASTI**, se observa que el problema del truncado ha sido superado (modificaciones en el proceso matemático) y la dispersión de valores ha disminuido (incremento en la selectividad de las ventanas espectrales). También se aprecia una mayor similitud en la clasificación cualitativa entre los valores de VRASTI corregido y RASTI.

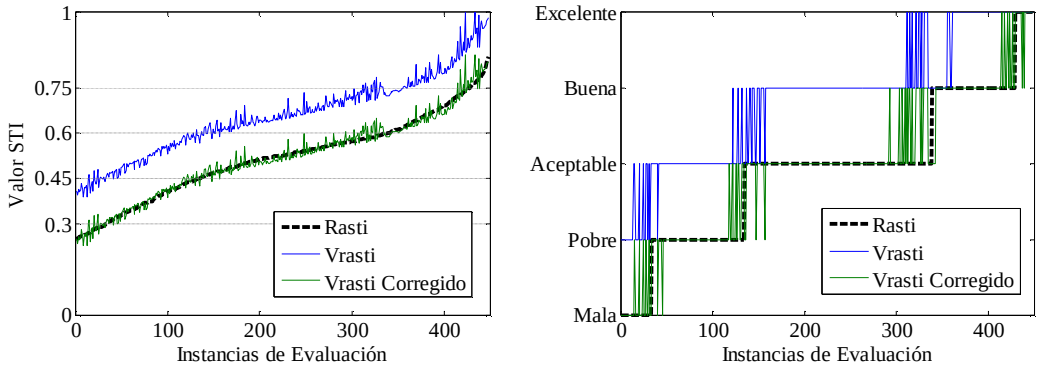

Figura 7. Comparación continua y discreta de resultados STI obtenidos con **VRASTI original** y con RASTI para un locutor particular.

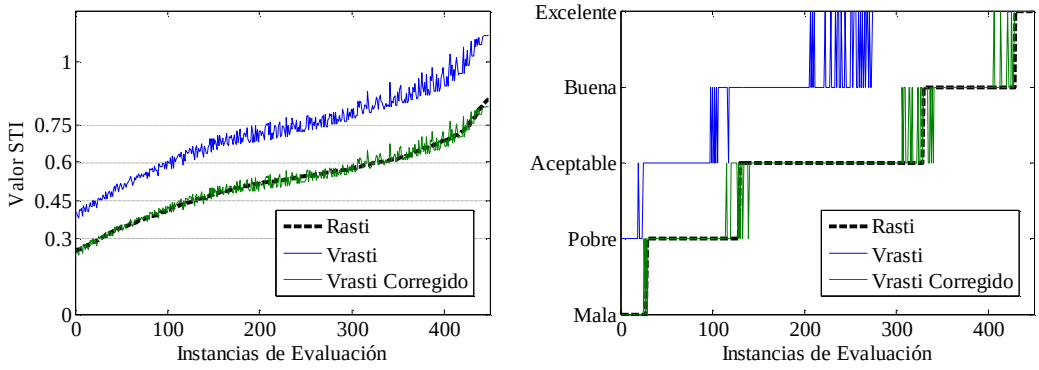

Figura 8. Comparación continua y discreta de resultados STI obtenidos con **VRASTI mejorado** y con RASTI para un locutor particular.

Se especifican en tabla 1 los **rangos** de niveles de ruido y reverberación para los que se ha probado el modelo presentado de **"VRASTI mejorado"**.

En la tabla 2 se muestran los errores de **"VRASTI mejorado"** con respecto a los de RASTI en idénticas condiciones de ensayo, para los dos tipos de locutor definidos. Se definieron 3 errores [7]: Medio (μ), Máximo y Probable, el cuál se obtiene sumando a la media del error el doble del desvío estándar. Este último equivale a un entorno en el que caerían el 95 % de las muestras en una distribución normal y se utiliza sólo como referencia.

Parámetro	Rango
SNR	[-30 30] dB
TR	[70 800] ms

Tabla 1. Rango de parámetros para los que fue probada la versión mejorada de VRASTI.

		Locutor	
		Particular	General
Error	Medio (μ)	0,025	0,041
	Máximo	0,159	0,207
	Probable ($\mu+2\times\sigma$)	0,063	0,108

Tabla 2. Error en los resultados de la versión mejorada de VRASTI relativo a los obtenidos con RASTI.

8 Conclusiones

A través de las simulaciones realizadas se encuentra que es factible diseñar un método de medición de inteligibilidad de la palabra utilizando la voz como señal de prueba, habilitando así la posibilidad de realizar mediciones en vivo, quedando pendiente la realización de pruebas en campo para completar la validación.

Como característica desfavorable se puede mencionar una mayor carga computacional con respecto a la implementación de RASTI ya que si bien se reemplaza el cálculo de un logaritmo sobre cada índice por la aplicación de un polinomio sobre la media de éstos, se implementan ventanas espectrales Hanning lo que implica un mayor procesamiento de datos. Además si se pretende adecuar por locutor, si bien es necesario hacerlo solo una vez por cada uno, esto requiere un elevado número de operaciones.

De todos modos, dada la potencia de los microprocesadores dedicados existentes hoy día, la implementación de la versión mejorada de VRASTI es totalmente factible.

Referencias

[1] Steeneken Herman J. M. ; Houtgast T. (1980). "A physical Method for Measuring Speech Transmission quality". Journal of the Acoustical Society of America (67) 1, 318-326.

[2] International Standard IEC 60268-16: Sound System Equipment – Part 16 (2003). "Objective rating of speech intelligibility by speech transmission index", International Electro technical Commission.

[3] Etter W. ; Moschytz G. S. (1994). "Noise reduction by noise adaptive spectral magnitude expansion". Journal Audio Eng. Soc. (42) 5.

[4] Steeneken Herman J. M. ; Houtgast T. (1985). "The Modulation Transfer Function in Room Acoustics". B & K Technical Review.

[5] Steeneken Herman J. M. ; Houtgast T. (1985). "RASTI: A tool for evaluating auditoria". B & K Technical Review.

[6] Miyara Federico. (1999). "Acústica y Sistemas de Sonido". UNR Editora, Rosario, Argentina.

[7] Camusso Javier H. ; Lértora Cristian E. (2008). “Medición de la Inteligibilidad de la Palabra utilizando Señales de Voz”. Proyecto Final de Carrera. Escuela de Ing. Electrónica, Univ. Nacional de Rosario, Argentina.