Naturaleza hablante: estrategias para generar voces creíbles a partir de elementos o fenómenos de la naturaleza

Speaking Nature: Strategies for Generating Credible Utterances of Nature Elements or Phenomena

Recibido: 9 de mayo de 2013
Aprobado: 20 de mayo de 2013

Andrea Lorena Aldana Blanco*

(Este trabajo de investigación fue financiado por el Grupo de Tecnologías Musicales –Music Technology Group - MTG– de la Universidad Pompeu Fabra de Barcelona, España. El documento original se realizó en el idioma inglés bajo el nombre: «Speaking Nature: Strategies for Generating Credible Utterances of Nature Elements or Phenomena»).

Resumen

Este documento explora la técnica de síntesis de voz *cross-synthesis*, a partir de sonidos de la naturaleza y grabaciones de voz con el fin de generar expresiones creíbles que preserven las características del sonido de la naturaleza escogido, en este caso el mar, conservando a su vez la inteligibilidad del mensaje. El mecanismo propuesto de producción sonora del mar se articula asociando olas a sílabas, y se proponen estrategias para controlar la velocidad y trayectoria de las olas. La automatización del proceso requiere el uso de descriptores tímbricos y el análisis fonético del texto que el mar debe recitar.

Palabras clave

Síntesis de voz, sonidos de la naturaleza, *cross-synthesis*, LPC, modelo *source-filter*, diseño sonoro, descriptores.

* Ingeniera de Sonido, Universidad de San Buenaventura, Bogotá. Magíster en Tecnologías del Sonido y de la Música, Universidad Pompeu Fabra, Barcelona, España. Profesora del programa de Ingeniería de Sonido, Universidad de San Buenaventura, Bogotá. E-mail: aaldana@usbbog.edu.co

Abstract

This paper explores the voice synthesis technique *cross-synthesis*, based on sounds of nature and voice recordings as a mean to create credible utterances that preserve the characteristics of the chosen sound of nature, in this particular case the sea, while maintaining the intelligibility of the message. The sound production mechanism proposed links syllables with waves. The automation of the process requires timbre descriptors and the phonetic transcription of the sea utterances.

Keywords

Voice synthesis, sounds of nature, cross-synthesis, LPC, source-filter model, sound design, descriptors..

I. Introducción

El creciente interés por el sonido como elemento fundamental en el cine, los videojuegos, la publicidad y los medios a nivel global, ha generado un aumento en el desarrollo de nuevas tecnologías y la exploración de diferentes campos del conocimiento con el fin de crear o recrear espacios sonoros «reales» y creíbles. El proceso de concepción de la idea sobre cómo debe sonar un espacio y la creación del mismo, se conoce como "diseño sonoro". Campos del conocimiento como la percepción humana y el procesamiento de señales, entre otros, son fundamentales para entender, crear e implementar un concepto de sonido para un determinado espacio.

Andy Farnell en su libro Designing Sound[1] describe el proceso de diseño sonoro como una estructura soportada por tres pilares del conocimiento, estos son: el físico, el matemático y el psicológico [1]. Lo anterior indica que este proceso debe tener en cuenta cómo las técnicas de grabación y procesamiento de la señal de audio aplicadas a un sonido afectan nuestra percepción de él y cómo estos procesos evocarán emociones relacionadas con el evento sonoro, además es necesario conocer y entender las propiedades físicas que constituyen un sonido para poder construirlo.

El proceso de diseño sonoro puede incluir varios elementos que van desde crear los efectos Foley de una película, es decir, grabar y sincronizar con el video aquellos sonidos que se producen por la interacción entre el personaje y la utilería en una escena determinada pero que se han perdido en la grabación del sonido directo durante el rodaje, hasta crear voces para personajes «no-humanos», por ejemplo voces de robot, como lo hizo Ben Burtt en la película WALL-E[2]. Finalmente, la meta es construir un espacio sonoro creíble. En varios casos el proceso de creación de voces no humanas está basado en las posibilidades que ofrecen los efectos digitales de audio (DAFX por su nombre en inglés *Digital Audio Effects*). Estos efectos, son técnicas de procesamiento de la señal

1 http://mitpress.mit.edu/books/designing-sound

2 http://www.disney.es/wall-e/

que permiten modificar un determinado sonido con el fin de producir una diferencia perceptual [2] y han sido clasificados de acuerdo con los distintos métodos que se aplican a la señal de entrada. En la literatura se han definido las siguientes categorías: Clasificación de acuerdo con el procesamiento de la señal [2] [3] [4] [5], clasificación perceptual [6], y clasificación basada en el tipo de control [7]. Sin embargo, en 2006 fue propuesto un método de clasificación interdisciplinario [8] que tiene en cuenta tanto las técnicas de procesamiento de la señal de audio como el resultado perceptual de los efectos, con el fin de facilitar la comunicación entre compositores, diseñadores sonoros, ingenieros DSP, musicólogos y demás profesionales involucrados en el campo de los efectos de audio.

Este trabajo de investigación se enfoca en explorar la aplicación de efectos de audio sobre una categoría de elementos sonoros que son ampliamente utilizados cuando se quiere crear o recrear un determinado espacio: los sonidos de la naturaleza. El concepto de los sonidos de la naturaleza en el proceso de diseño sonoro ha sido abordado principalmente desde la perspectiva del paisaje sonoro, sin embargo, estos sonidos también han sido grabados y transformados de distintas maneras para mejorar y enriquecer escenas de películas, videojuegos, instalaciones artísticas, experiencias 4D de parques temáticos, música, entre otros.

Este proyecto explora métodos de síntesis de voz enfocados en sonidos de la naturaleza y grabaciones de voz con el fin de generar expresiones creíbles que preserven las características del sonido o fenómeno de la naturaleza escogido manteniendo la inteligibilidad del mensaje, partiendo de la siguiente pregunta de investigación: ¿Si pudiera, como hablaría el mar?

II. Marco teórico

A continuación se explican algunos conceptos que fueron importantes en el desarrollo de esta investigación:

1. El concepto de diseño sonoro

El diseño es un proceso que reúne varios campos del conocimiento y en el que distintas herramientas son aplicadas con el fin de construir un espacio sonoro. La complejidad de este proceso se basa en la necesidad del diseñador sonoro de entender y dominar distintos campos como lo son el tecnológico, la percepción humana, la estética y la semiótica [9].

2. Procesamiento de la voz

El procesamiento de la señal de voz involucra categorías como: el reconocimiento de voz, la identificación del hablante, la codificación de la voz y la síntesis de la voz, entre otros. Para profundizar en las técnicas de procesamiento de voz, es necesario entender el proceso de producción del habla.

2.1 Producción del habla

La producción de habla comienza en el momento en el que el cerebro envía la información de lo que queremos decir e involucra cuatro etapas: iniciación, fonación, proceso oro-nasal y articulación [10].

El proceso de iniciación se presenta cuando el aire es expulsado de los pulmones, después comienza el proceso de fonación cuando el aire pasa a través de la laringe e interactúa con las cuerdas vocales y la glotis. Si la glotis tiene una apertura angosta, la combinación de la presión del aire de los pulmones con la tensión de las cuerdas vocales genera una vibración de las mismas y produce un sonido periódico (sonoro). Por el contrario, si la glotis tiene una apertura amplia, el aire de los pulmones pasa a través de ella y la fricción genera un sonido sordo. Una vez el aire ha pasado, viajará a través de las cavidades nasales y orales. Finalmente, el proceso de articulación ocurrirá en la boca de acuerdo con las posiciones de distintos articuladores como: los dientes, la lengua y los labios superior e inferior.

2.2 El modelo *source-filter* (fuente-filtro)

En la producción del habla, el modelo *source-filter* involucra la combinación de dos aspectos, el primero de ellos son las cuerdas vocales actuando como la señal de excitación y el segundo son las cavidades de la boca y la nariz actuando como un resonador. El resonador es equivalente a un filtro, enfatiza ciertas frecuencias (formantes) y atenúa otras (antiformantes).

Se puede observar el comportamiento del sistema fuente-filtro observando el espectro. En efecto, en el caso de sonidos sonoros podemos decir que los harmónicos muestrean el timbre, indican el valor de amplitud y fase del timbre, de manera que para sonidos graves se tiene mejor resolución frecuencial que para sonidos agudos. Es importante notar que aquí definimos el timbre como la envolvente espectral, una versión suavizada del espectro de amplitud que recorre los picos (harmónicos). En cambio, los harmónicos en sí son parte de la excitación, muestran la naturaleza periódica de la excitación. Por ejemplo, el primer harmónico tiene una frecuencia igual al inverso del periodo de vibración de la glotis, y los siguientes harmónicos tienen frecuencias múltiples a ella.

Para aplicar el modelo mencionado, es necesario realizar dos pasos: primero debemos estimar la envolvente espectral, y en segundo lugar, necesitamos separar la señal de excitación y el filtro para realizar una transformación de uno de ellos. Una vez hayamos realizado el proceso, hacemos una combinación de tipo *source-filter*.

Para estimar la envolvente espectral, podemos utilizar varias técnicas:

- Vocoder de canal: esta técnica se basa en bandas de frecuencia, ya que calcula los valores RMS (*Root Mean Square – valor cuadrático medio*) por cada banda para estimar la envolvente espectral.

- Predicción lineal: esta técnica estima un filtro *all-pole* que representa el contenido espectral de un sonido. Cuando el orden de este filtro es bajo, solo los formantes son tomados, obteniendo así la envolvente espectral.
- Cepstrum: la técnica cepstrum realiza un suavizado del logaritmo del espectro de la FFT (*Fast Fourier Transform – Transformada Rápida de Fourier*), en decibeles, con el fin de separar la parte que varía lentamente, es decir, su envolvente espectral y la parte que varía rápidamente,o sea, la señal de excitación [2].

2.3 Transformaciones de tipo *source-filter: Cross-synthesis*

El efecto de audio *cross-synthesis* toma dos señales de entrada y genera una tercera a partir de la combinación de las dos primeras. La idea central se basa en «imprimir» la envolvente espectral del segundo sonido en el primero y a la vez preservar la altura del primer sonido. Para esto, es necesario extraer una envolvente espectral que varía en función del tiempo (segundo sonido) y aplicarla a una señal (primer sonido). Dentro de esta transformación también es utilizada una técnica conocida como «*Whitening*» que consiste en remover la envolvente espectral del primer sonido antes de aplicar a esta señal la del segundo.

III. Metodología

El sonido de la naturaleza escogido para implementar el modelo de *cross-synthesis*, fue el sonido del mar. Inicialmente, observamos los elementos del mar que sirven como base para describir su comportamiento, y de esta manera, establecer un concepto sobre cómo hablaría.

1. El concepto del mar

El sonido producido por las olas lleva un mensaje que explica el estado del mar en un determinado lugar y momento. Estas olas no son constantes en términos de la velocidad, la fuerza, la dirección y la continuidad y así mismo estos cambios dependen de las condiciones climáticas, las condiciones físicas del lugar en el que se ha formado el mar, etc.

El estado emocional del habla es parte de la prosodia, y está dado por el ritmo, la acentuación y la entonación. La duración de la sílaba, la altura, y la intensidad son factores que generan emoción en el habla. Si realizamos una analogía entre prosodia en el habla y prosodia en el mar, partiendo desde el punto de vista del comportamiento de las olas, entonces el largo de estas, la intensidad y la continuidad con la que llegan a la costa pueden dar el punto de partida del concepto de un mar que habla.

Al considerar que las sílabas son grupos de fonemas que dan un ritmo natural en el habla[3], y que en el mar este ritmo está dado por las olas, se puede pensar en emparejar cada ola a una sílaba, ya que esto nos permite variar el volumen, la velocidad y observar como estos afectan la calidad y la inteligibilidad del mensaje.

3 Applied Speech and Audio Processing, p. 40.

Es importante comentar que esta investigación se realizó en el idioma inglés y esto se ve representado en las frases escogidas. En otros idiomas, elementos como el número de fonemas y la información que los formantes y la altura transmiten pueden ser muy diferentes. En el idioma inglés, la inteligibilidad del mensaje incrementa cuando el segundo y tercer formante (F2 y F3) están presentes. El tono en sí mismo no contribuye mucho a la inteligibilidad del mensaje, sin embargo, en lenguajes tonales como el mandarín, la inteligibilidad del mensaje depende del tono. La tabla 1 muestra la fragmentación por sílabas de la frase «*Once upon a time* (había una vez)».

Word	Syllable 1	Syllable 2
Once	Once	
Upon	U	Pon
A	A	
Time	Time	

Tabla 1. Fragmentación por sílabas de la frase «Once upon a time»

2. Construyendo la base de datos de olas y grabaciones de voz

2.1 Grabaciones de voz

Las frases escogidas para el proyecto fueron frases famosas de distintas películas o cuentos infantiles. La tabla 2 muestra algunas de las frases utilizadas en el desarrollo de la investigación..

Frase	Película/Año
Bond, James Bond	Dr No. (1962)
Houston we have a problem	Apollo 13 (1995)
Once upon a time	
Once upon a time, there was a little girl named Goldilocks	Ricitos de oro

Tabla 2. Frases grabadas

Las grabaciones de la voz se hicieron en modo susurro porque de esta forma desde la grabación nos podemos acercar más al sonido producido por el mar, ya que si modelamos el mar como ruido blanco filtrado entonces los susurros tendrán características similares. En cambio, si grabamos palabras pronunciadas normalmente, los fonemas sonoros tendrán un espectro harmónico que será muy diferente de ruido blanco filtrado. Las grabaciones fueron realizadas con frecuencia de muestreo de 44100 muestras/segundo y cuantización de 16 bits en formato WAV.

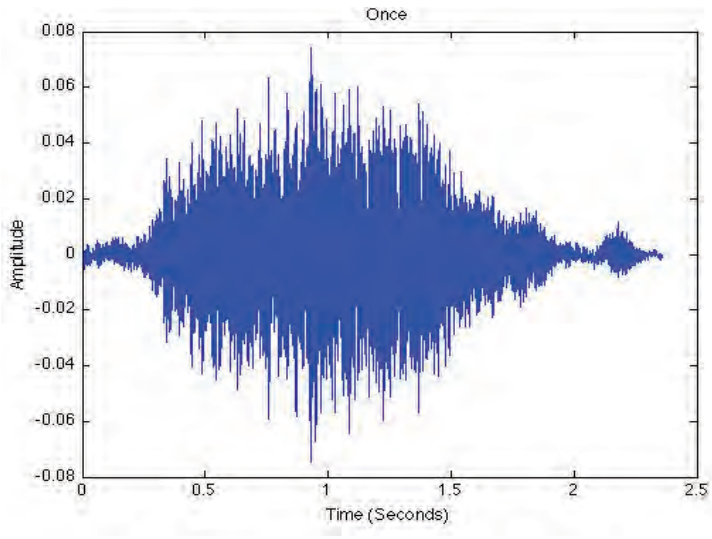

Figura 1. Forma de onda de la sílaba «Once» susurrada.

La comparación de las envolventes espectrales de una grabación del mar, la grabación de la palabra «once» susurrada y la misma palabra producida de manera normal, se presentan en la figura 2.

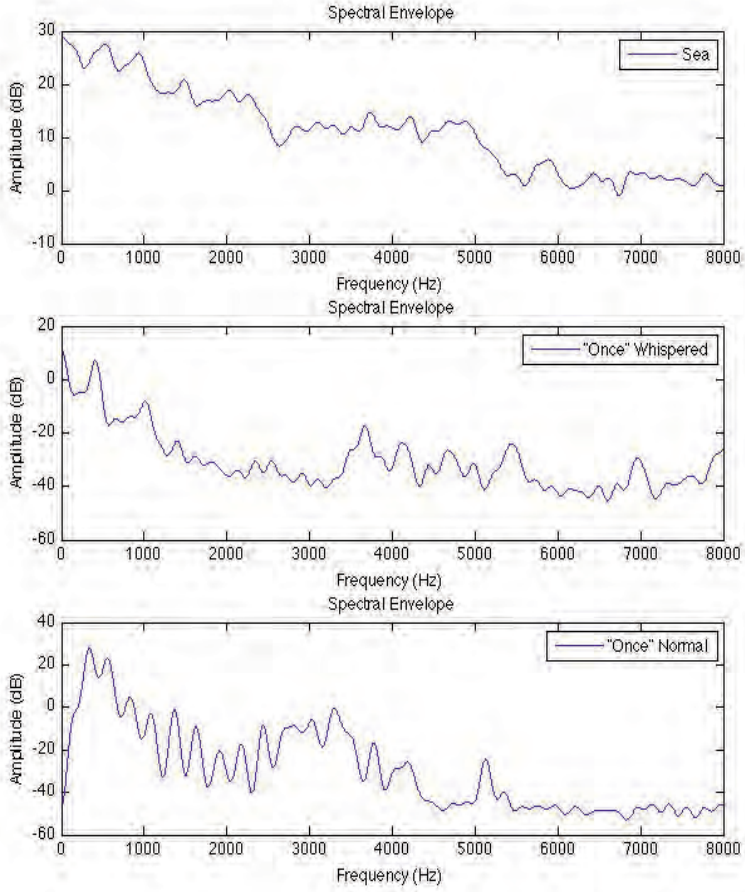

Figura 2. Comparación de las envolventes espectrales: mar (primera gráfica), susurro «once» (segunda gráfica), y «once» sin susurrar (tercera gráfica). En el eje *Y* se puede observar la amplitud y en el eje *X* la frecuencia.

En las grabaciones de voz se utilizó el micrófono M-Audio Nova. La siguiente tabla muestra las especificaciones técnicas del micrófono:

M-Audio NOVA***** 
Condensador
Patrón polar: Cardioide
Respuesta en frecuencia: 20 Hz to 18000 Hz
Sensibilidad: 16 mV/Pa (-36dBV)
Impedancia: 200 ohms

Tabla 3. Especificaciones técnicas del micrófono M-Audio Nova.

Inicialmente las grabaciones de voz se realizaron con el micrófono dinámico Shure C606, sin embargo, estas primeras pruebas mostraron que al tener que realizar las grabaciones en forma de susurro era necesario contar con un micrófono que tuviera mayor sensibilidad, por esto se escogió finalmente el M-Audio Nova.

2.2 Grabaciones del mar

Las grabaciones del mar se realizaron utilizando la grabadora portátil Edirol R1 y una pantalla anti-viento construida con espuma. Estas grabaciones se hicieron en las playas de la ciudad de Barcelona, algunas fueron tomadas en la costa y otras en los rompeolas. La base de datos de olas se construyó a lo largo del año durante el tiempo seco de las estaciones de invierno, primavera y verano. Se buscó tomar las muestras en días que tuvieran poco viento con el fin de reducir el ruido en las olas grabadas. Todas las grabaciones fueron realizadas en estéreo con frecuencia de muestreo de 44100 muestras/segundo y cuantización de 16 bits en formato WAV.

3. Caracterización de olas y sílabas

3.1 Caracterización de las olas

Con el fin de caracterizar las olas utilizamos descriptores de audio, ya que estas eran diferentes en términos de la duración, el volumen y el tono. Para obtener los descriptores utilizamos el MIR Toolbox[5].

Como punto de referencia para caracterizar las olas se tomaron los descriptores de audio correspondientes a la energía global, el centroide y el brillo de la señal en cuatro segmentos de cada ola. Estos segmentos se establecieron de la siguiente manera:

4 http://www.m-audio.com/products/en_us/Nova.html

5 https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

- Segmento 1: comienzo del audio (Punto 1) hasta el comienzo de la ola (Punto 2)
- Segmento 2: comienzo de la ola (Punto 2) hasta el punto de ruptura de la ola (Punto 3)
- Segmento 3: punto de ruptura de la ola (Punto 3) hasta el punto de posruptura de la ola (Punto 4)
- Segmento 4: punto de posruptura de la ola (Punto 4) hasta el final de la ola (Punto 5)

Los tiempos correspondientes a cada segmento fueron tomados manualmente al escuchar el comportamiento de cada ola. Esta información fue guardada en un archivo de texto para ser utilizada en la obtención de los descriptores. En la siguiente tabla se incluyen los valores por segmento obtenidos para el descriptor brillo en algunas olas: un valor igual a cero es un sonido muy opaco, un valor igual a uno es un sonido muy brillante.

Brillo				
Número de la ola	Segmento 1	Segmento 2	Segmento 3	Segmento 4
Ola 1	0.4722	0.5429	0.5483	0.5556
Ola 2	0.5116	0.5223	0.5607	0.4627
Ola 3	0.5173	0.4746	0.6005	0.7074
Ola 4	0.5443	0.4766	0.6004	0.6885

Tabla 4. Valores del descriptor brillo

3.2 Caracterización de las sílabas

Para la caracterización de las sílabas utilizamos el transcriptor fonético Unisyn Lexicon desarrollado por Susan Fitt del Centro para la Investigación de Tecnologías del Habla (Centre for Speech Technology Research[6]) de la Universidad de Edinburgo, con el fin de descomponer cada frase en palabras, sílabas y fonemas. Esto nos permitió establecer los tipos de fonemas y acentos presentes en cada frase. Al ingresar la frase «Houston we have a problem» al transcriptor fonético obtuvimos la siguiente transcripción:

/ Houston we have a problem
/ [Sil] [« h j u . s t @ n] [w i] [« h { v] [@] [« p r A . b 5 @ m] [Sil]
Sil h j u s t @ n ws i h { v @ p r A b 5 @ m Sil

La transcripción fonética nos permitió obtener información sobre las sílabas acentuadas, ya que estas se representaban con el símbolo ´"´. A partir de esta información, creamos un vector de ganancia en el cual las sílabas con acento tenían un mayor valor. Adicionalmente, generamos un vector de tiempos dependiendo de la composición de las palabras, de esta forma, si una palabra tenía dos o más sílabas, estas estarían más cerca en términos del tiempo que palabras monosílabas. Esto nos permitió crear un mar que hablara más rápido o más lento.

6 http://www.cstr.ed.ac.uk/projects/unisyn/

4. Procesamiento *source-filter*

Con el fin de generar la síntesis de voz a partir de las olas y las grabaciones de voz, realizamos dos pasos. El primero consistió en estimar la envolvente espectral y el segundo en realizar una separación fuente-filtro para transformar una de las señales y posteriormente combinarlas. La envolvente espectral la obtuvimos utilizando la técnica cepstrum. El cepstrum es la transformada inversa de Fourier del logaritmo del espectro de una señal. A partir de esta técnica podemos separar la fuente y el filtro de una señal.

Figura 2. Proceso de obtención del cepstrum de una señal.

Una vez estimada la envolvente espectral, realizamos el proceso de *cross-synthesis*, aplicando la envolvente espectral del sonido de la voz al sonido del mar en un rango de frecuencias determinado.

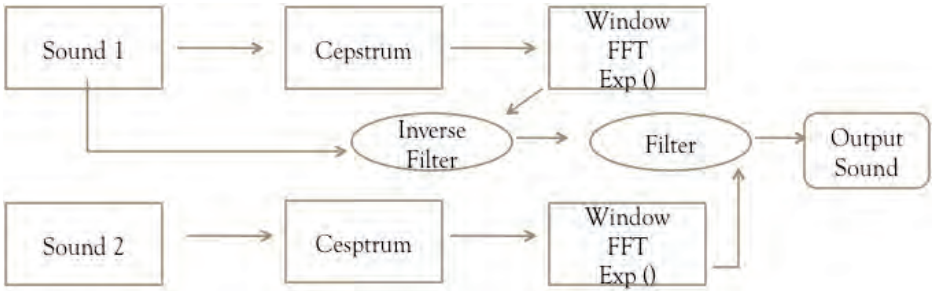

Figura 3. Proceso de síntesis de voz a partir del método de cross-synthesis.

La figura 4 muestra las envolventes espectrales de la fuente y del filtro en un tiempo determinado.

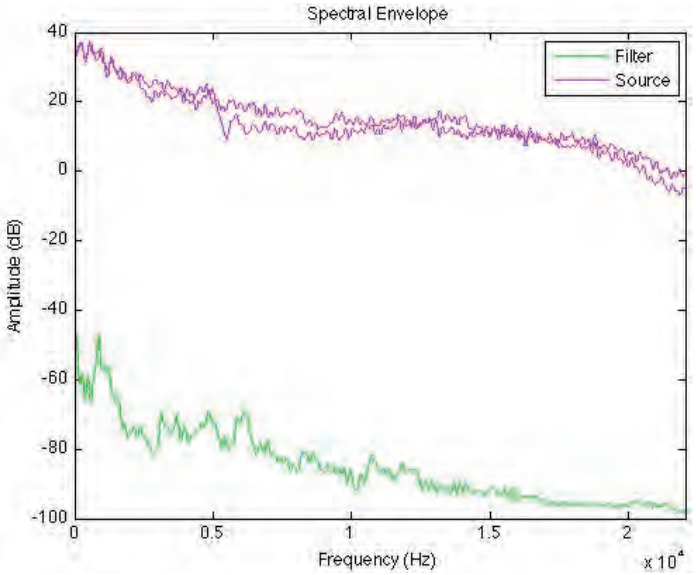

Figura 4. Envolvente espectral de la fuente y el filtro. Filtro (filter), línea verde. Fuente (Source), línea rosada.

Una vez obtenidas las envolventes espectrales se realizó una normalización ya que las amplitudes de las envolventes de la voz y del mar pueden ser muy distintas. Por esta razón, en el dominio de la frecuencia, promediamos las envolventes espectrales de los dos sonidos para estimar la amplitud del filtro que aplicaríamos a la fuente. Sin embargo, al hacer esto de forma lineal encontramos que para ciertas bandas de frecuencia relacionadas con el rango de la voz, la amplitud del mar decae rápidamente y arrastra la amplitud de los formantes, lo cual genera una pérdida de inteligibilidad. Por esta razón, implementamos una regresión polinomial a las envolventes de la voz y del mar, lo que nos permitió determinar un filtro que preserva el comportamiento general de la envolvente espectral del mar. Este proceso lo realizamos en dB.

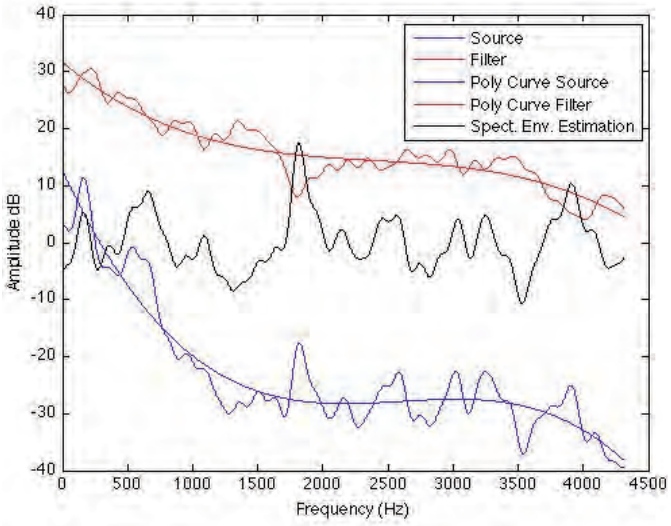

Figura 5. Estimación de la envolvente espectral. Regresión Polinomial. Fuente (*Source*) y estimación polinomial de la envolvente de la fuente (*poly curve source*), línea azul. Filtro (*Filter*) y estimación polinomial de la envolvente del filtro (*poly curve filter*), línea roja. Estimación de la envolvente espectral, línea negra.

Una vez hecha la normalización y estimadas las envolventes espectrales de los dos sonidos, aplicamos la envolvente del filtro a la envolvente de la fuente solo en un rango de frecuencias. El rango de frecuencias escogido fue entre 300 Hz y 4000 Hz. Fuera de ese rango el filtro tiene un valor de 0 dB.

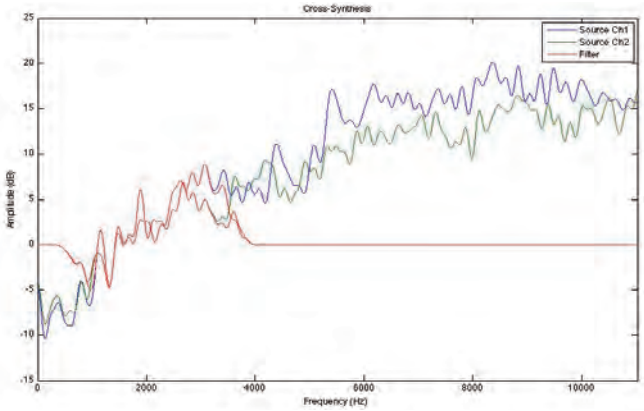

Figura 6. *Cross-synthesis* (600 Hz - 4000 Hz). Fuente canal 1 (*Source* ch1), línea azul. Fuente canal 2 (*Source* ch2), línea verde. Filtro (*filter*), línea roja.

El filtro aplicado al sonido de la fuente no estuvo solo limitado a un rango de frecuencias, sino también a un rango de tiempo. Para establecer los rangos de tiempo utilizamos los puntos mencionados en la sección 3.1 (Caracterización de las olas). El proceso de *cross-synthesis* se realizó entre el punto 1 de la ola hasta el punto 3, es decir, la posruptura de la ola. Para conseguir este efecto, se realizó un escalamiento en el tiempo de la grabación de voz con respecto a los puntos de tiempo de procesamiento predefinidos de la ola. Una vez realizado este proceso notamos que escalar la duración de las consonantes no sonaba natural, por esta razón, decidimos alargar en mayor proporción las vocales. Esto lo hicimos tomando manualmente los tiempos de inicio y final de las vocales en cada sílaba y almacenándolos en un archivo de texto para su posterior procesamiento.

5. Mezcla de sonidos

Después de generadas las olas hablantes, creamos el sonido ambiente a partir de olas individuales. La ganancia de cada sílaba estuvo dada por la acentuación obtenida a partir de la transcripción fonética, por otra parte, la ganancia del ambiente se estableció de forma manual.

En el caso del paneo, las olas que contenían sílabas se ubicaron en el centro. Para los ambientes se generaron dos capas, una de estas paneada a la izquierda y la otra capa paneada a la derecha [rango: 0 (izquierda) a 1(derecha)].

IV. Evaluación y análisis de resultados

Generadas las frases del mar hablante, se realizó un test de escucha subjetivo de tipo MOS (*Mean opinion Score*), en el cual se evaluaron aspectos como la inteligibilidad del mensaje, la calidad del sonido y si el sonido sintetizado preservaba las características del sonido del mar.

Puntaje	Descripción
5	Excelente
4	Buena
3	Regular
2	Pobre
1	Mala

Tabla 5. Escala *Mean Opinion Score* (MOS)

Las frases con las cuales se realizó la evaluación se presentan en la tabla 6. La primera parte del test consistió en escuchar la voz sintetizada sin conocer la frase real bajo la cual se construyó el mar hablante. En la segunda parte se les mostró la frase a los participantes.

Frase
Once upon a time
Houston we have a problem
And they lived happily ever after
Once upon a time there was a girl named Goldilocks

Tabla 6. Frases utilizadas en la evaluación

Los resultados de la encuesta subjetiva se presentan en la tabla 7. El aspecto A corresponde a la calidad del sonido y el aspecto B a la inteligibilidad del mensaje.

Frase	Primera parte del test	Aspecto A	Aspecto B	Segunda parte del test	Aspecto A	Aspecto B
Once upon a time		3,07	1,69		2,69	2,15
Houston we have a problem		2,46	1,92		2,69	1,76
And they lived happily ever after		2,69	2		2,46	1,61
Once upon a time there was a girl named Goldilocks		2,84	2,3		3	1,53

Tabla 7. Resultados de la evaluación subjetiva

El siguiente gráfico muestra los resultados de la pregunta: ¿El sonido sintetizado, preserva las características del mar?

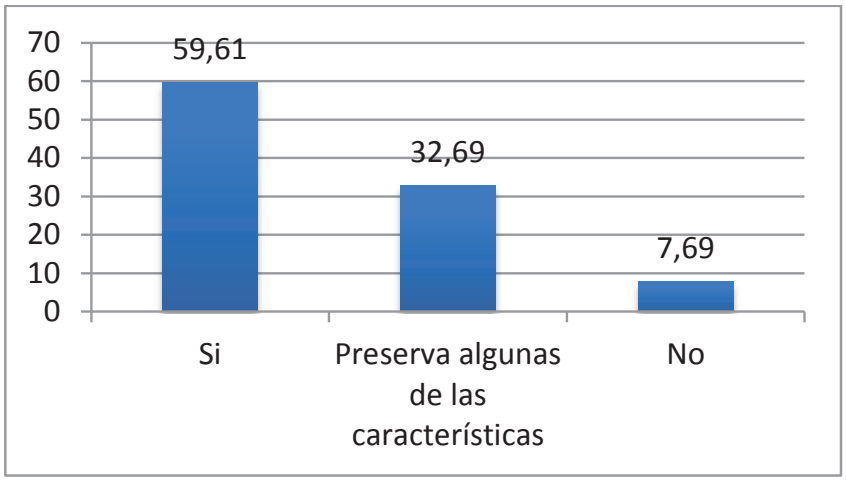

Gráfico 1. Resultados de la evaluación subjetiva. Valores en porcentaje %.

Los resultados de la encuesta subjetiva permiten concluir que la inteligibilidad del mensaje disminuyó con el aumento del número de palabras y consonantes explosivas en una frase, además la tabla 7 muestra que la calidad del sonido tuvo puntajes más altos en la escala MOS que la inteligibilidad del mensaje. Sin embargo, el gráfico 1 muestra que el 59,61% de los encuestados considera que el concepto de mar hablante propuesto en este proyecto, preserva las características del mar. Una vez presentada la frase a los participantes del test en la segunda etapa de la prueba, la calificación en la escala MOS

aumentó en la frase con menor número de sílabas *(«once upon a time»)* y presentó una disminución para las demás frases que contenían mayor número de sílabas y consonantes explosivas.

V. Conclusiones

- La aproximación al concepto de un mar que habla fue considerado principalmente como una representación creíble de las características de mar.
- La inteligibilidad de un fonema incrementa o decrece en relación con el rango de frecuencias seleccionado para el filtro. Las vocales tienden a tener formantes más bajos, y por lo tanto, el filtro puede funcionar bien al tener una frecuencia de corte media. En el caso de las consonantes explosivas la frecuencia de corte debe ser más alta. Esto puede ser mejorado al implementar una banda de frecuencia activa que varíe dependiendo del tipo de fonema en cada trama, el punto de partida sería la transcripción fonética.
- La inteligibilidad del mensaje es altamente dependiente de las consonantes y del número de palabras que componen la frase sintetizada.
- Las vocales son más fáciles de entender por su alto contenido de energía.
- La técnica de *cross-synthesis* proporciona un amplio rango de posibilidades para desarrollar el modelo de una mar que habla, sin embargo, es necesario trabajar aún más en la síntesis de las consonantes, especialmente las explosivas, con el fin de mejorar la inteligibilidad a nivel general.

Referencias

[1] A. Farnell, *Designing Sound,* 2010.

[2] U. Zölzer, *DAFX: Digital Audio Effects*. J. Wiley & Sons, 2002.

[3] S. Orfanidis, *Introduction to Signal Processing*. Prentice Hall Int. Editions, 1996.

[4] C. Roads, *The Computer Music Tutorial*. Cambridge, Massachusetts: MIT Press, 1996.

[5] F. Richard Moore, *Elements of Computer Music*. Englewood Cliffs, N.J.: Prentice Hall,1990

[6] X. Amatriain, J. Bonada, A. Loscos, J. L. Arcos, and V. Verfaille, *Content-based transformations*, J. New Music Research, vol. 32, no. 1, pp. 95–114, 2003.

[7] V. Verfaille, Marcelo M. Wanderley and P. Depalle. «Mapping strategies for gestural control of adaptive digital audio effects». J. New Music Research, vol. 35, no. 1, pp. 71-93, 2006.

[8] V. Verfaille, C. Guastavino and C. Traube. *An interdisciplinary approach to to audio effect classification*. 9th Int. Conf. Digital Audio Effects (DAFx-06), Montreal, Canada, pp. 107-13, 2006.

[9] M. Liljedahl, J.Fagerlönn, *Methods for sound design: a review and implications for research and practice*. Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound, NY, USA, 2010.

[10] http://www.ugr.es/~ftsaez/fonetica/production_speech.pdf