Análisis y Tratamiento del Habla Esofágica para la Mejora de su Comprensión

ROCÍO SESMA ALCALDE, JORGE MIQUÉLEZ ECHEGARAY Y YOLANDA BLANCO RODRÍGUEZ
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA
ETSII Y IT, UNIVERSIDAD PÚBLICA DE NAVARRA
CAMPUS DE ARROSADÍA S/N, 31006 PAMPLONA
Correo electrónico: *jomiquel@gmail.com*

ABSTRACT: This paper presents an empirical study of the esophageal speech and proposes a method to improve its intelligibility using speech synthesis. In the analysis was found that the esophageal speech is characterized by a low fundamental frequency, while the formant patterns are similar to those of normal speakers. In the implemented method to improve the intelligibility, the treatment is different for voiced and unvoiced frames of the signal. While the unvoiced frames are preserved the voiced frames are re-synthesized using linear prediction. Among the studied source signals, the polynomial one is used due to the naturalness of the synthesized sounds and the fundamental frequency is raised up to normal values, keeping its variability, resulting in an intelligible and more natural speech.

1. Introducción

El tratamiento del cáncer de laringe puede hacer necesaria la extirpación de la laringe. Esto alterará el mecanismo de producción del habla, ya que el aire de los pulmones se escapa a través del estoma abierto en el cuello de los pacientes y se han eliminado las cuerdas vocales que son las encargadas de sonorizar el aire. Estas personas deben aprender a llenar de aire la parte alta de su esófago y a usar como vibrador el orificio de entrada al mismo.

El estudio del habla que se ha llevado a cabo se basa en el modelo del habla propuesto en [1], ya que permite separar el concepto de fuente vocal (ligado a la vibración del aire y a las cuerdas vocales) del de tracto vocal (ligado a la cavidad resonante formada por boca, lengua, paladar y labios). El hecho de extirpar las cuerdas vocales afectaría a la forma de la fuente, pero no a la del tracto.

2. Análisis de la voz esofágica

El estudio se ha realizado a partir de distintas señales grabadas a un varón adulto operado de laringe, y digitalizadas con una frecuencia de muestreo de 22 KHz, y una codificación PCM con 16 bits por muestra.

En el primer análisis tiempo-frecuencia se observa que, frente a espectrogramas de habla normal, las formantes no difieren en absoluto, mientras que el pitch es muy poco estable, y de una frecuencia muy baja, entre 50 y 70 Hz (ver Fig. 1). Esto hace que la voz suene ronca y ruidosa.

¿Qué le sucede a la fuente vocal de una persona laringectomizada? Si se realiza un análisis de la frecuencia fundamental del habla esofágica surge un problema, y es que los métodos tradicionales de detección del pitch [2] no dan unos resultados satisfactorios, debido a la falta de periodicidad del habla. Para evitar esto, el método que se ha utilizado parte de la Señal de Error de Predicción [3], de la que se calcula la señal de covarianza de pequeños segmentos solapados. La forma de onda así obtenida tiene mayor periodicidad que la señal de voz de partida, y está sincronizada con ella.

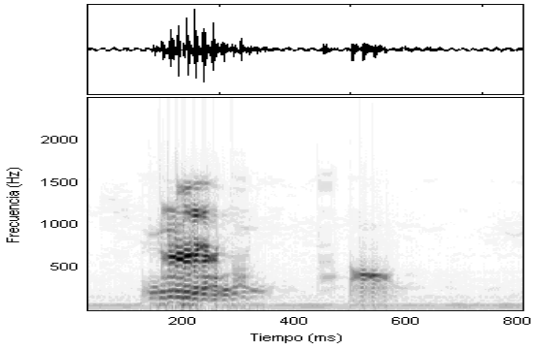

Fig. 1. Forma de onda y espectrograma de la palabra "martes" pronunciada por un laringectomizado.

3. Elección de la fuente de voz

Suponiendo que el filtro del tracto vocal no se altera tras la operación, se podría mejorar el habla esofágica mediante un proceso de análisis, para extraer las características del filtro que modela el tracto vocal, y síntesis, filtrando una fuente vocal optimizada.

Se han probado distintos tipos de fuentes de voz para reemplazar la fuente existente en vocales aisladas: fuente monopulso, suma de sinusoides y modelo de fuente polinomial.

3.1. Fuente monopulso

La fuente más sencilla es la excitación monopulso [1], que se caracteriza por un único pulso al comienzo del período. Esta fuente genera un habla inteligible, pero con un sonido metálico muy poco natural.

3.2. Suma de sinusoides

Otra alternativa como fuente es la suma de funciones senoidales de la frecuencia fundamental y sus armónicos hasta un margen de frecuencia considerado [4]. Se han estudiado dos casos: uno, con todas las sinusoides en fase, y otro en el cual, el k - ésimo armónico de la frecuencia fundamental tiene una fase $\phi_k = \frac{\pi \cdot (k-1)^2}{L}$ siendo L el número total de armónicos.

Si la fase de todas las ondas es la misma al comienzo de cada período, se obtiene una fuente

bastante parecida al monopulso, debido a que unos senos se anulan con los otros. Los resultados no mejoran mucho.

Si las sinusoides están desfasadas, la fuente obtenida tiene una apariencia ruidosa, manteniendo una amplitud similar a lo largo del período. El sonido logrado es bastante mejor que en los casos anteriores.

3.3. Modelo de fuente polinomial

La última opción estudiada como fuente de voz es un modelo polinomial de tercer orden [5]. En este caso, la forma de onda resultante se acerca más al proceso natural del habla, teniendo en cuenta parámetros como los tiempos de apertura y cierre de la glotis (Fig. 2). La síntesis obtenida con esta fuente resulta al oído la más natural de todas las comentadas.

4. Método utilizado para la mejora del habla

En las grabaciones realizadas, lo primero que se observaba era una interferencia sinusoidal de 50 Hz, posiblemente derivada de la red eléctrica. Para cancelar esta interferencia, es necesario utilizar una técnica de filtrado adaptativo [6], debido a que las frecuencias fundamentales del habla esofágica están en torno a los 50 - 70 Hz. Si se filtrara la señal con un filtro de rechazo de banda, la información del pitch se perdería.

Una vez eliminada la interferencia, se separan los segmentos sonoros de la señal de voz de los sordos, ya que el tratamiento es distinto, según el tipo de trama de que se trate. Se clasifica en función de los cruces por cero y de la desviación estándar de la señal.

Los segmentos sordos se mantienen en la señal sintética igual que en la original.

En las tramas sonoras hay que sustituir la fuente vocal original defectuosa por la sintética, que deberá tener la frecuencia que correspondería a un sujeto sano, pero manteniendo la evolución de la frecuencia que sigue el paciente. Entonces, se realiza un análisis síncrono con el pitch. Se toman ventanas de 30 ms, de las cuales se estima la frecuencia fundamental de cada período de la señal, como se explica en la sección 2 de este artículo. Sincronizándose con el inicio del período, y tomando una ventana de igual duración que el mismo, se calculan los coeficientes de predicción lineal (LPC), que caracterizan el filtro del tracto vocal en ese tramo de señal. Con estos LPC se filtran períodos de fuente vocal.

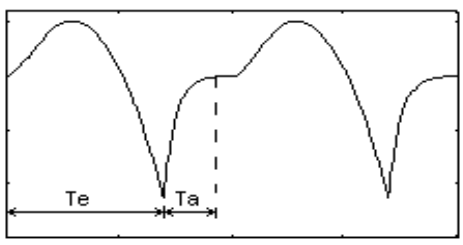

Fig. 2. Dos períodos del modelo de fuente polinomial, donde se indica Te y Ta, los tiempos de apertura y cierre de la glotis.

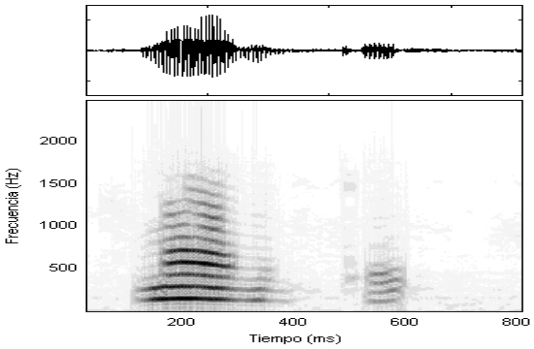

Fig. 3. Forma de onda y espectrograma de la síntesis de la palabra "martes" de la Fig. 1.

Una vez realizado este proceso en los tramos sonoros y obtenidos sus correspondientes tramos sintéticos, éstos se concatenan con los tramos sordos de la señal original. De esta forma se reconstruye el discurso inicial.

Un ejemplo del resultado de este proceso puede verse en la Fig. 3.

5. Conclusiones

Con el método propuesto se consiguen síntesis de gran naturalidad y una buena comprensión. Además, deja una puerta abierta a la búsqueda de algoritmos y equipos que puedan realizar la conversión voz esofágica - voz normal en tiempo real.

Sería conveniente ampliar el estudio realizado aquí con mayor número de grabaciones, procedentes de distintas personas, y tanto hombres como mujeres.

La calidad de grabación también puede ser mejorada, con el uso de equipos con mayores prestaciones y en ambientes más robustos ante el ruido.

Agradecimientos

Los autores de esta investigación quieren agradecer a Javier Sánchez Uzcarré su disponibilidad y la ayuda prestada, así como al Dr. Javier Medina, jefe de la sección de otorrinolaringología del Hospital Virgen del Camino, de Pamplona, su interés y apoyo.

Referencias

1 Oppenheim, A. y Schafer, R. (1989) *Discrete – Time Signal Processing*, 815-821 (Prentice-Hall)

2 Rabiner L., Cheng, J. Rosenberg A. y McGonegal C. "A Comparative Performance Study of Several Pitch Detection Algorithms" *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-24**, 399-404, Oct. 1976.

3 Rabiner, L., y Schafer, R. (1978*) Digital Processing of Speech Signals*, 421-424 (Prentice-Hall)

4 Van Santen, J., Sproat, R., Olive, J., y Hirschberg, J. (1996) *Progress in Speech Synthesis*, 57-70 (Springer)

5 Van Santen, J., Sproat, R., Olive, J., y Hirschberg, J. (1996) *Progress in Speech Synthesis*, 27-39 (Springer)

6 Widrow, B. y Stearns, S. (1985) *Adaptative Signal Processing*, 302-367 (Prentice-Hall)