APLICACIÓN DE MÉTODOS DE SUSTRACCIÓN ESPECTRAL PARA REDUCCIÓN DE RUIDO

Martín Haimovich[1] y Ignacio Veiga[2]

[1]Universidad Nacional de Tres de Febrero
mhaimov1995@gmail.com
[2]Universidad Nacional de Tres de Febrero
ignacioveigagiusti@gmail.com

Resumen —

1. INTRODUCCIÓN

En el presente trabajo se desarrollará un algoritmo de reducción de ruido para aplicación en señales sonoras. Mediante el mismo, se busca poder eliminar el ruido en fragmentos de la señal en los que se esperaría que haya silencio, ya sea porque no hay voz o música o porque no hay información inteligible o separable del ruido de fondo. Para el mismo, se han consultado desarrollos previos y tomado la decisión de aplicar determinados algoritmos en busca de una mejora en los resultados del proceso. Se ha de notar la dificultad que presenta, en general, la separación completa de una señal y del ruido sumado a la misma cuando no se tienen ambas componentes por separado. Al tener sólo el resultado, se debe tomar algún criterio para caracterizar al ruido, así como diversas decisiones en cuanto a la forma de aplicar filtros que puedan reducirlo efectivamente sin eliminar o afectar a la parte de la señal que se quiere conservar. Incluso utilizando los algoritmos más complejos y eficaces, si el ruido es de muy alto nivel en relación a la señal que se desea separar del mismo, o si ambas componentes tienen un contenido frecuencial similar, muy probablemente la reducción del mismo no sea satisfactoria. Dado esto, es menester la definición de objetivos claros con respecto al alcance del algoritmo a desarrollar, entendiendo sus limitaciones y comportamiento general.

2. MARCO TEÓRICO

Para realizar este proyecto se investigó acerca de los distintos métodos de reduccion de ruido existentes, encontrando como el primero en importancia al algoritmo desarrollado por Boll en su trabajo "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", el cual plantea la problemática de que al ser una voz con ruido de fondo igual a la suma de la señal limpia y el ruido residual, este último no se puede eliminar por completo pero si hacer una reducción considerable para que la señal final sea aproximadamente la señal limpia, que es lo que generalmente se quiere escuchar.

Previamente antes de ese trabajo, las formas de reducir el ruido de fondo de una señal eran mediante micrófonos cancela ruido, modificaciones internas de los algoritmos del procesador de voz, para compensar explícitamente la contaminación de voz, y mediante pre procesadores reductores de ruido. Los micrófonos tenían la limitación de que sólo generaban una reducción en el ruido de frecuencia inferior a 1 kHz, mientras que las modificaciones internas de los algoritmos requeridas no se podían llevar a cabo en ese entonces por falta de dinero, tiempo y esfuerzo, y los pre procesadores existentes generaban una independencia de la eliminación del ruido con respecto a las implementaciones de los procesadores de voz.

Los objetivos del trabajo de Boll fueron desarrollar una técnica de eliminación de ruido, implementar un algoritmo de calculo eficiente y testear su performance en ambientes ruidosos. Esto se desarrolló restándole a la voz ruidosa el ruido, para obtener una estimación de la magnitud del espectro en frecuencia de la voz limpia. Para esto se necesita estimar el ruido promedio medido durante la actividad sin voz. Boll utiliza este enfoque, en el cual se puede aproximar el error de aproximacion espectral y se

desarrollan los métodos secundarios para reducirlo.

2.1. Análisis de eliminación de ruido sustractiva:

Se obtiene una estimación espectral del ruido midiendo la señal cuando no hay voz presente, para luego restar esta estimación a la señal de voz ruidosa y así determinar la parte limpia de la señal. Mientras el valor esperado del espectro del ruido de fondo sea igual al mismo cuando hay actividad vocal, el entorno de ruido permanecerá estacionario. Si el entorno cambia a un nuevo estado estacionario, hay un limite de tiempo, de alrededor de 300 ms, para estimar una nueva magnitud del espectro del ruido de fondo, antes de que se comienze a escuchar la voz. Finalmente, se asume que la reducción de ruido sólo es posible removiéndolo de la magnitud del espectro de la señal sucia.

2.2. Modelo de suma de ruido:

Se puede calcular la suma de una señal de ruido ventaneada n(k) con una señal de voz ventaneada s(k) dando como resultado la siguiente expresión:

$$x(k) = s(k) + n(k) \tag{1}$$

Luego, haciendo las respectivas transformadas de Fourier:

$$X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega}) \tag{2}$$

Donde X es la transformada de la señal de voz ruidosa x(k) la cual se puede descomponer en el producto entre su magnitud y su fase:

$$X(e^{j\omega}) = |X(e^{j\omega})| * e^{j\theta_X} \tag{3}$$

2.3. Estimador de resta espectral:

Se despeja de (2) la señal limpia, quedando igualada a la diferencia entre el espectro de la señal ruidosa y el del ruido:

$$S(e^{j\omega}) = X(e^{j\omega}) - N(e^{j\omega}) \tag{4}$$

Sabiendo que:

$$N(e^{j\omega}) = |N(e^{j\omega})| * e^{j\theta_N} \tag{5}$$

Se reemplaza a la magnitud del espectro de ruido por el valor medio del mismo medido cuando no hay voz presente mientras que su fase es reemplazada por la de la señal de voz ruidosa, quedando (4):

$$S_1(e^{j\omega}) = (|X(e^{j\omega})| - |\mu(e^{j\omega})|) * e^{j\theta_X} \tag{6}$$

Siendo

$$\mu(e^{j\omega}) = E|N(e^{j\omega})| \tag{7}$$

2.4. Error espectral:

El error espectral resultante de este estimador viene dado por:

$$\epsilon(e^{j\omega}) = S_1(e^{j\omega}) - S(e^{j\omega}) = N(e^{j\omega}) - \mu(e^{j\omega}) * e^{j\theta_X} \tag{8}$$

Este puede ser atenuado mediante:

2.4.1. Promedio de la magnitud

Teniendo en cuenta (8), se puede usar el promedio local de la magnitud del espectro de la señal de voz ruidosa para lograr una atenuación del error, reemplazando la magnitud de esta señal por su conjugado, siendo este:

$$|\bar{X}(e^{j\omega})| = 1/M * \sum_{i=0}^{M-1} |X_i(e^{j\omega})| \tag{9}$$

Siendo $X_i(e^{j\omega})$ la iésima ventana transformada de $x(k)$ Quedando:

$$S_A(e^{j\omega}) = [|\bar{X}(e^{j\omega})| - |\mu(e^{j\omega})|] * e^{j\theta_X} \tag{10}$$

El motivo por el cual se promedia es que el error espectral se puede comenzar a aproximar:

$$\epsilon(e^{j\omega}) = S_A(e^{j\omega}) - S(e^{j\omega}) \cong |\bar{N}| - \mu \tag{11}$$

donde

$$|\bar{N}(e^{j\omega})| = 1/M * \sum_{i=0}^{M-1} |N_i(e^{j\omega})| \tag{12}$$

Por lo tanto, el promedio de $|N(e^{j\omega})|$ va a converger a $\mu(e^{j\omega})$ a medida que se toma un promedio con mayor cantidad de valores. El problema de este método es que como la voz humana no es estacionaria, se debe promediar a un tiempo limitado, ya que si el promediado es mayor al de 3 ventanas con un tiempo de duración de 38.4 ms, se disminuye considerablemente la inteligibilidad de la señal. Pero la mayor desventaja del promediado de la magnitud es que pueden llegar a aparecer sonidos transistorios de corta duración no deseados.

2.4.2. Rectificación de media onda

Para cada frecuencia, se define que si la magnitud del espectro de la señal de voz ruidosa es menor que la del ruido promediado μ, entonces se define el valor de cero a la salida. El estimador se define como:

$$S_1(e^{j\omega}) = H_R(e^{j\omega}) * X(e^{j\omega}) \tag{13}$$

Siendo

$$H_R(e^{j\omega}) = \frac{H(e^{j\omega}) + |H(e^{j\omega})|}{2} \tag{14}$$

$$H(e^{j\omega}) = 1 - \frac{\mu(e^{j\omega})}{|X(e^{j\omega})|} \tag{15}$$

La ventaja de la rectificacion de media onda es que reduce el ruido por μ y cualquier variación de los tonos del ruido es eliminada, pero el problema aparece cuando la suma del ruido con la voz a una frecuencia determinada es menor a μ, ya que la información de la voz en esa frecuencia termina siendo eliminada generando asi una disminución de la inteligibilidad.

2.4.3. Reducción de ruido residual

Luego de la rectificación, la suma de la voz y el ruido permanece por encima del promedio del espectro del ruido μ. En la ausencia de voz, la diferencia $N_R = N - \mu * e^{j\theta_n}$ es conocida como ruido residual, y tendrá una magnitud entre cero y un valor máximo medido durante la ausencia de la voz. Al transformar la señal de nuevo al dominio del tiempo, el ruido residual sonará como la sumatoria de generadores de tonos con frecuencias fundamentales aleatorias las cuales aparecerán y desaparecerán en intervalos de tiempo de aproximadamente 20 ms. Durante la presencia de la voz, el ruido residual será percibido en las frecuencias que no son enmascaradas por la voz. Los efectos audibles del ruido residual se pueden reducir tomando ventaja de este ventana a ventana aleatoriamente. Específicamente a un intervalo de frecuencia dado, el ruido residual variará de forma aleatoria en amplitud en cada ventana de análisis, este puede ser eliminado reemplazando su valor por su mínimo valor elegido mediante el análisis por ventanas. Se puede tomar el mínimo valor cuando la magnitud de la estimación rectificada H_1 definida en (13) es menor al ruido residual máximo calculado durante la ausencia de la voz. Este reemplazo depende de la amplitud que tenga S_1:

1. Si la amplitud de S_1 está por debajo del máximo ruido residual, y este varía radicalmente en el análisis ventana a ventana, entonces hay una gran posibilidad de que el espectro, a esa frecuencia, sea generado por el ruido, por lo tanto se lo eliminará tomando el mínimo.

2. Si S_1 esta por debajo del máximo ruido residual, pero tiene un valor constante cercano a este, hay altas posibilidades de que el espectro a esa frecuencia sea debido a la voz con un bajo nivel de energía, por lo tanto, tomando el mínimo ruido se podrá retener la información necesaria.

3. Si S_1 está por encima del máximo ruido residual, es porque hay voz en esa frecuencia, por lo tanto, es suficiente con remover su sesgo.

El esquema de la reducción de ruido residual se define mediante:

$$|S_1 i(e^{j\omega})| = |S_1 i(e^{j\omega})| \tag{16}$$

cuando $|S_1 i(e^{j\omega})| \geq max|N_R(e^{j\omega})|$

Mientras que

$$|S_1 i(e^{j\omega})| = min S_1 j(e^{j\omega})(j = i-1, i, i+1) \tag{17}$$

cuando $|S_1 i(e^{j\omega})| < max|N_R(e^{j\omega})|$

Siendo

$$S_1 i(e^{j\omega}) = H_R(e^{j\omega}) * X_i(e^{j\omega}) \tag{18}$$

$max|N_R(e^{j\omega})| =$ máximo valor del ruido residual medido durante la ausencia de la voz

Pese a que el nivel de reducción de ruido es similar al del método de promediado sobre las 3 ventanas, la desventaja es que se necesita una mayor energía para almacenar el máximo ruido residual y los valores de magnitud de las tres ventanas adyacentes.

2.4.4. Atenuación adicional de la señal cuando no hay voz presente

El contenido de la energia del estimador de la resta espectral S_1 con respecto al promedio del ruido durante la ausencia de la voz μ brinda una información precisa de la presencia de la voz en la señal, dentro de un análisis por ventanas determinado. Esto se determina mediante:

$$T = 20 \log \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{S_1(e^{j\omega})}{\mu(e^{j\omega})}\, d\omega \tag{19}$$

Cuando T es menor a -12 dB, quiere decir que no hay voz y la única información disponible es la del ruido residual que sobrevivió al rectificador y a la selección del mínimo valor. En este caso se pueden realizar 3 acciones:

1. No hacer nada: esto implica una amplificación del ruido durante la ausencia de voz en la señal.

2. Definir la salida igual a cero: similar a no hacer nada, pero en los momentos en los que hay voz presente.

3. Atenuar la salida por un factor fijo, logrando una disminución del nivel de la señal equivalente a -30 dB

Esta medición se detalla de la siguiente forma:

$$S_1(e^{j\omega}) = S_1(e^{j\omega}) \tag{20}$$

si $T \geq -12dB$. Mientras que

$$S_1(e^{j\omega}) = cX(e^{j\omega}) \tag{21}$$

siendo $20 \log c = -30dB$

2.5. Implementación del algoritmo

2.5.1. Ventaneo con ventana de Hanning

La señal pasada por un conversor AD es segmentada y ventaneada con un solapamiento de un 50 % y cada segmento es multiplicado por una ventana de Hanning. Luego sumando cada secuencia ventaneada se obtiene una

copia de la secuencia original. Se define un largo de ventana mayor al doble del período de tono máximo esperado para una resolución de frecuencia adecuada. Por ejemplo, para una frecuencia de muestreo de 8 kHz, se usa un largo de ventana de 256 muestras desplazadas en pasos de 128 muestras.

2.5.2. FFT

Se realiza la transformada rápida de Fourier en cada parte de la señal ventaneada. El tamaño de la FFT debe ser igual al de la ventana (en el ejemplo anterior, debería ser de 256 muestras)

2.5.3. Promedio de la magnitud de la señal transformada

La varianza de la estimación del ruido espectral es reducida promediando tantas magnitudes espectrales como sea posible. Sin embargo, como la voz no es estacionaria, el intervalo total de tiempo disponible para el promediado local es limitado. La cantidad de promedios posibles está limitado por el numero de ventanas que se pueden introducir en el intervalo de tiempo en el que la voz es estacionaria. La elección del ancho de ventana y del intervalo promediado debe comprometerse entre requisitos en conflicto. Para tener una resolución espectral aceptable, se necesita un largo de ventana mayor al doble del período tonal más largo esperado. Para una varianza mínima del ruido, se requiere un largo numero de ancho de ventana para promediar. Mientras que para una resolución temporal aceptable, se requiere una banda de intervalo estrecha. Un compromiso razonable entre reducción de varianza y resolución temporal puede ser de 3 promediados. Esto da como resultado un intervalo de tiempo de análisis efectivo de 38 ms.

2.5.4. Resta del sesgo

El método de resta espectral requiere una estimación del valor esperado del ruido a cada intervalo de frecuencia

$$\mu_N = E|N| \tag{22}$$

Esta estimación se hace promediando la magnitud del espectro de la señal cuando no hay voz presente. Estimando μ_N de esta forma, se generan ciertas restricciones cuando se implementa el método. Si el ruido permanece estacionario cuando la voz esta presente, entonces será necesario un período de calibración inicial de solo ruido. Se puede calcular una estimación de μ_N en este período (de aproximadamente un tercio de segundo). Mientras que si el ruido no permanece estacionario, entonces se deberá calcular una nueva estimación de μ_N cada vez que el espectro del ruido cambie, debiéndose hacer antes de la eliminación del sesgo. Dado que la estimación es calculada usando la señal de ruido durante la ausencia de voz, es necesario un cambio de voz. Cuando no hay voz, se puede recalcular el promedio del espectro del ruido. Si la magnitud del espectro del ruido se modifica más rápido que su estimación, no se podrá usar el promediado en tiempo para estimar el ruido. De todas formas, si el valor esperado del espectro del ruido cambia luego de haber calculado su estimación, entonces la reducción del ruido mediante la eliminación del sesgo será menos efectiva, pudiendo llegar a ser dañina, es decir, se puede llegar a eliminar la voz donde hay un pequeño ruido presente.

2.5.5. Rectificación de media onda

La estimación de la resta espectral se obtiene restando la magnitud del espectro del ruido μ de la de la señal $|X|$. De este modo:

$$|S_1(k)| = |X(k)| - \mu(k) = H(k) * X(k) \tag{23}$$

Siendo k=0,1,...,L-1, mientras que L es la longitud de la DFT y H está definida por:

$$H(k) = 1 - \frac{\mu(k)}{|X(k)|} \tag{24}$$

Una vez hecha la resta, todos los valores que den negativos se definen como ceros. Estas diferencias negativas representan frecuencias donde la suma de la voz con el ruido local es menor al ruido esperado.

2.5.6. Reducción del ruido residual

Como ya se dijo, el ruido que permanece luego de remover la media se puede eliminar seleccionando el valor mínimo de la magnitud de las 3 ventanas adyacentes de análisis en cada intervalo de frecuencia, en donde la amplitud es menor que el máximo ruido residual medido durante la ausencia de la voz. Dado que el mínimo se elige entre los valores que hay dentro de las ventanas temporales, la modificación incluye un retraso de una ventana. Pese a esto, esta mejora es considerada superior a la del promediado de tres ventanas, debido a que una cantidad equivalente de ruido suprimido resultó sin los efectos adversos del suavizado espectral de alta energía.

2.5.7. Cálculo del detector de actividad de la voz y atenuación del ruido adicional durante la señal sin voz

La última mejora en la reducción de ruido consiste en eliminar la señal cuando hay ausencia de la voz. Como ya se planteó, se debe mantener un equilibrio entre la magnitud y características del ruido percibido durante la presencia de voz y el ruido percibido durante la ausencia de esta. Se define un detector de voz usando un generador espectral. Este detector necesita tener un umbral que defina la ausencia de la voz. Este umbral se establece en aproximadamente -12 dB y es definido para asegurar que únicamente sean atenuadas las señales que tengan como información solamente ruido de fondo.

2.5.8. Transformada rápida inversa de Fourier IFFT

Luego de la eliminación del sesgo, la rectificación, la reducción del ruido residual y la eliminación de la señal cuando la voz está ausente, se procede a reconstruir la

onda en el dominio del tiempo a partir de la magnitud modificada de la ventana central.Dado que solo se genera informacion real, se calculan dos ventanas temporales en simultáneo usando la IFFT. La información de las ventanas se superponen y se van agregando para formar la secuencia de salida de la voz.

2.6. Limitaciones del modelo de Boll

El modelo de Boll encuentra ciertas limitaciones a la hora de realizar la resta espectral. Una de sus más importantes limitaciones es producida por el llamado ruido musical", el cual se genera durante la rectificación de media onda, cuando se definen como cero a los valores de la señal que son menores al ruido estimado. Estos valores generan pequeños picos en frecuencias aleatorias del espectro convirtiéndose, en el dominio temporal, en tonos de muy poca duracion y cuyas frecuencias varían aleatoriamente de ventana a ventana. Además, aparecen y desaparecen en intervalos temporales de 20 a 30 ms. Estos tonos conforman el ruido musical y es una cuestión a mejorar. Otra limitacion del modelo es el uso de la fase del ruido que produce un empeoramiento de la calidad de la voz sintetizada. No es nada fácil estimar la fase de la voz limpia y además genera una complejidad mucho mayor en la mejora del algoritmo. Pero pese a que no es lo mejor para la calidad de la señal procesada, es mas importante atacar el ruido musical mencionado anteriormente. Es por eso que en este trabajo, el objetivo será disminuir el ruido musical que se genera con el método de sustracción espectral de Boll.

2.7. Resta espectral multibanda

En esta mejora del algoritmo, el espectro de la voz está dividido en N bandas superpuestas y la resta espectral se realiza en cada banda de forma independiente. El proceso de dividir la señal de la voz en diferentes bandas puede hacerse tanto usando filtros pasa bandas en el dominio del tiempo como usando ventanas apropiadas en el dominio de la frecuencia. La estimación del espectro de la voz limpia en la iésima banda se obtiene mediante:

$$|X_e i(\omega_k)|^2 = |Y_i(\omega_k)|^2 - \alpha_i \delta_i |D_i(\omega_k)|^2 \qquad (25)$$

estando ω_k limitado entre $b_i < \omega_k < e_i$. Donde $\omega_k = \frac{2\pi k}{N}$ (siendo $k = 0, 1, ..., N-1$) son las frecuencias discretas, $|D_e i(\omega_k)|^2$ es la potencia del ruido estimado durante la ausencia de la voz, α_i es el factor de la resta de la iésima banda y δ_i es una banda adicional. Mientras que b_i y e_i son las frecuencias inicial y final de la iésima banda de frecuencia. Haciendo la resta de esta forma, se puede definir en cada banda de frecuencia el proceso, haciendo al algoritmo más personalizado y evitando eliminar infomracion necesaria.

3. PROCEDIMIENTO

El procedimiento experimental se realiza a través de un programa desarrollado utilizando el lenguaje de programación Python. La señal a analizar se importa al programa utilizando la biblioteca SoundFile, debiendo tener en consideración la frecuencia de muestreo de la misma para su correcta vectorización y posterior análisis.

En primera instancia, se define un algoritmo para el reconocimiento del ruido a filtrar. Mediante los valores de los parámetros temporales Short Time Energy y Zero Crossing Rate se puede desarrollar un método de reconocimiento de fragmentos en los que la señal sólo contenga ruido. Se considera que no hay actividad vocal en intervalos en los que haya un alto nivel de energía por frame y un bajo ZCR. La complicación será definir los umbrales para ambos valores, pero se puede considerar una primera aproximación por defecto tomando el valor promedio de cada uno de los parámetros y utilizando ese valor como umbral. En casos donde haya poca actividad vocal, esto probablemente detecte muchos picos del ruido como actividad vocal, en dichos casos será necesario la definición de un criterio diferente. Una vez obtenidos los fragmentos de ruido se pueden obtener las características del ruido de fondo, su fase y la estimación de su espectro.

3.1. Implementación del método de Boll

Una vez logrado el reconocimiento y caracterización del ruido a filtrar, se aplica el método de Boll para reducción de ruido. La señal es segmentada y ventaneada con un solapamiento de un 50 % y cada segmento es multiplicado por una ventana de Hanning. Se define un largo de ventana mayor al doble del período de tono máximo esperado para una resolucion de frecuencia adecuada. Se toma 5 kHz como el tono máximo de la voz con aporte significativo de energía, si bien se reconoce que en otras circunstancias la energía en altas frecuencias podría ser importante para el procesamiento de voces [1].

Luego de ventanear la señal, se realizará la FFT a cada fragmento ventaneado de la señal utilizando las funciones de la biblioteca SciPy, para luego realizar el promedio de la magnitud como se detalla en la ecuación (9).Esto se utilizará junto con el promedio estimado del ruido para realizar la rectificación de media onda, detallada en la ecuación (15).

Se procede luego con el proceso de reducción de ruido residual que persista en las frecuencias no enmascaradas por la voz, finalizando con la atenuación de la señal en los intervalos sin voz, reduciendola por un factor fijo definido para lograr una reducción de 30 dB.

Por último se efectúa la transformada inversa de Fourier para así obtener la señal con ruido reducido.

3.2. Implementación del método de sustracción espectral multibanda

Como alternativa superadora al método de Boll, se propone mejorar la detección de fragmentos a filtrar mediante el análisis separado en bandas de frecuencias. Para ello, se debe modificar el método a partir de la obtención de la amplitud del espectro de la señal. Al espectro obtenido, se le aplicarán ventanas de Hanning de ancho W_k y solapamiento de 50 % para separar las bandas de frecuencia.También será necesario hacer lo mismo para la estimación del ruido, obteniendo sus valores por banda de frecuencia a aplicar en el cálculo para cada banda correspondiente. Los valores α_i y δ_i se definirán según la bibliografía consultada [2] como:

$$\alpha_i = \begin{cases} 4{,}75 & SNR_i < -5 \\ 3/20(SNR_i) & -5 < SNR_i < 20 \\ 1 & SNR_i > 20 \end{cases} \tag{26}$$

$$\delta_i = \begin{cases} 1 & f_i < 1 \text{ kHz} \\ 2{,}5 & 1 \text{ kHz} < f_i < (f_s/2) - 2 \text{ kHz} \\ 1{,}5 & f_i > (f_s/2) - 2 \text{ kHz} \end{cases} \tag{27}$$

4. RESULTADOS Y ANÁLISIS

5. CONCLUSIONES

REFERENCIAS

[1] Andrew J. Lotto Brian B Monson Eric J. Hunter y Brad H. Story. «The perceptual significance of high-frequency energy in the human voice». En: *Frontiers in Psychology* 5:587 (2014).

[2] Shashikant L. Sahare Anuradha R. Fukane. «Different Approaches of Spectral Subtraction method for Enhancing the Speech Signal in Noisy Environments». En: *International Journal of Scientific Engineering Research* 2.5 (2011).