Revista Cubana de Ciencias Informáticas
ISSN: 1994-1536
ISSN: 2227-1899
Editorial Ediciones Futuro

Reyes Díaz, Flavio J.; Roble Gutiérrez, Alejandro;
Hernández Sierra, Gabriel; Calvo de Lara, José Ramón
Filtrado wiener para la reducción de ruido en la verificación de locutores.
Revista Cubana de Ciencias Informáticas, vol. 12, núm. 3, 2018, Julio-Septiembre, pp. 152-162
Editorial Ediciones Futuro

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 3, Julio-Septiembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 152-162
http://rcci.uci.cu

Tipo de artículo: Artículos originales
Temática: Reconocimiento de patrones
Recibido: 30/10/2017 | Aceptado: 06/06/2018

Filtrado wiener para la reducción de ruido en la verificación de locutores

Wiener filtering to noise reduction for speaker verification

Flavio J. Reyes Díaz, Alejandro Roble Gutiérrez, Gabriel Hernández Sierra, José Ramón Calvo de Lara

Centro de Aplicaciones de Tecnologías de Avanzada(CENATAV). 7a.A # 21406 e/ 214 y 216, Playa, La Habana, C.P. 12200, Cuba. Email: freyes,arobles,gsierra,jcalvo@cenatav.co.cu

*Autor para correspondencia: freyes@cenatav.co.cu

Resumen
Las señales de audio, incluida la voz, de alguna forma están expuestas al deterioro de su calidad debido a la incorporación de ruidos ambientales. Estos ruidos existentes en la señales de audio, provocan una degradación de la calidad en la información acústica del locutor, trayendo consigo una disminución de la eficacia en el reconocimiento de locutores. En este trabajo se realiza un análisis del comportamiento de algunos de los principales métodos de reducción de ruido: Filtro de Wiener y Sustracción Espectral, ante señales de voces ruidosas. Finalmente, se propone aplicar el filtrado de Wiener a la etapa de pre-procesamiento de las señales de un sistema de reconocimiento de locutores. La evaluación de nuestra propuesta se realizó sobre muestras telefónicas de la base de voces NIST SRE-08, con diferentes tipos de ruidos ambientales, obteniendo una mejora relativa del EER de un 4,94 % y 12,5 % para ambas condiciones de evaluación.

Palabras claves: filtro de Wiener, ruido, verificación de locutores

Abstract
Audio signals, including voice, are in some way exposed to quality deterioration due to the environmental noise incorporation. These noises existing in the signals, provoke a quality degradation in the speaker acoustic information, bringing with it a decrease of the speaker recognition performance. In this work an analysis of the behavior of some of the main noise reduction methods is performed such as: Wiener Filter and Spectral Subtraction, to pre-procesing the noisy signals. Finally, it is proposed to apply the Wiener filtering to the stage of pre-processing the signals of a speaker recognition system. Then, our proposal are evaluated on telephone session from NIST SRE-08 for different environmental noises types, obtaining an EER improvement of 4,94 % and 12,5 % for both evaluation conditions.

Keywords: *noise, speaker verification, Wiener filter.*

Introducción

Todas las señales de audio, incluida la voz, de alguna forma están expuestas al deterioro de su calidad, dado que en cualquiera de las etapas por las que puede pasar, emisión, propagación, captura, transmisión, almacenamiento y reproducción se puede introducir ruido. Se denomina ruido a toda señal no deseada que se mezcla con la señal deseada, en este caso, la voz. Las señales de voz son afectadas por diferentes tipos de ruido, como se describen en (Scheffer et al., 2013): el ruido ambiental, la distorsión propia del teléfono, los ruidos propios del canal por donde se transmite la voz, los ruidos de cuantificación y codificación, entre otros.

El reconocimiento automático de locutores (RAL) no se encuentra ajeno a esta problemática, debido a que el ruido aditivo provoca una degradación de la calidad de la información acústica del locutores existente en la señal de audio, provocando una disminución de la eficacia del RAL (Ming et al., 2007; Mandasari et al., 2012; Rajan et al., 2013). Hasta la actualidad se han propuesto disímiles métodos para reducir el efecto del ruido aditivo en el RAL, principalmente en las etapas de: procesamiento de las señales de audio y extracción de rasgos acústicos robustos.

Para enfrentar la degradación del audio, en la etapa de procesamiento de las señales se han aplicado 3 métodos de filtrado principalmente para reducir el efecto del ruido aditivo:

- la Sustracción Espectral (Davis, 2002): es uno de los primeros algoritmos de filtrados propuestos para cancelar el ruido aditivo de la señal ruidosa. Asumiendo que el ruido de una señal de voz es aditivo, este algoritmo sustrae el espectro de ruido del espectro de la voz y solo debe quedar el espectro de voz limpio. Para esto realiza una estimación del espectro de ruido en una región de la señal en que no haya voz y considerar que ese espectro no cambia a lo largo de la señal. Es muy efectivo para reducir la relación señal-ruido (SNR), si se logra estimar adecuadamente el espectro de ruido, pero introduce en la señal un nuevo ruido, conocido como ruido musical; el cual afecta la inteligibilidad en la señal de audio.
- el método $RASTA_{LP}$ (Boril et al., 2011): es un filtro paso bajo que se aplica a la señal ruidosa para reducir el efecto del ruido aditivo y la reverberación.
- el Filtro de Wiener (Agarwal and Cheng, 1999): presenta como principal objetivo minimizar el error medio cuadrático entre la señal de voz limpia y la ruidosa. Para alcanzar su objetivo se apoya en métodos estadísticos y reduce el ruido presente en la señal de audio corrupta de tal modo que la señal de salida del filtro se aproxime lo más posible a la señal deseada.

Para enfrentar la distorsión de la información acústica del locutor debido al ruido aditivo se han diseñado diferentes rasgos acústicos, específicamente para robustecer el RAL ante la variabilidad debido al ruido (Scheffer

et al., 2013): el cepstrum de la Modulación de la duración media (MDMC) (Mitra et al., 2012), los Coeficientes Cepstrales con Normalización de la Potencia (PNCC) (Kim and Stern, 2016) y los Coeficientes de la Envolvente de Hilbert (MHEC) (Sadjadi and Hansen, 2011).

Partiendo de los resultados prometedores alcanzados por las Redes Neuronales profundas (DNN) en la rama del reconocimiento automática del habla, se han realizado estudios para aplicar las DNN al RAL. En (McLaren et al., 2014; Richardson et al., 2016) se han propuesto métodos donde se aplica esta técnica para reducir el efecto del ruido sobre la señal de voz, aplicado pricipalmente sobre el espacio de los rasgos acústicos. Otros trabajos como (Pekhovsky et al., 2016; Plchot et al., 2016) proponen aplicar los autoencoders apilados formando DNN para mejorar la calidad de la señal de voz. Estos métodos reciben una señal corrupta por ruido y como resultado obtiene una señal limpia, por lo que los denominan Denoising DNN (por su definición en el Ingles).

A partir del estudio realizado en esta área de investigación, se detectaron diferentes problemas e inconvenientes para aplicar varios de los métodos antes mencionados. En el caso de los rasgos robustos, su principal problema es que fueron diseñados principalmente para enfrentar el ruido, por tanto, cuando son aplicados sobre escenarios donde las señales de voz no están corruptas, el RAL disminuye su eficacia. Por otra parte, las DNN requieren de grandes bases de voces para su correcto entrenamiento.

Por otra parte, el Filtro de Wiener fue uno de los métodos más utilizados para reducir el efecto del ruido y obtuvo los mejores resultados (Saedi et al., 2013; Ferrer et al., 2013) en la evaluación bianual NIST[1] SRE-2012 (Greenberg et al., 2013), donde por primera vez utilizan señales de evaluación afectadas por diferentes tipos de ruidos.

Teniendo en cuenta las inconvenientes antes mencionadas que presentan las DNN y los rasgos robustos, y basándonos en los resultados alcanzados por el Filtro de Wiener en la evaluación NIST SRE-2012, se propone aplicar el Filtro de Wiener para reducir el ruido en las señales de voz y con esto aumentar la eficacia del RAL sobre escenarios donde el ruido es muy variable.

Materiales y métodos

Los sistemas de reconocimiento de locutores que representan el estado del presentan diferentes etapas: el pre-procesamiento de las señales de audio, la extracción de rasgos acústicos, el cálculo de i-vector, la compensación de la variabilidad de sesión y el cálculo de la puntuación de la similitud. A continuación se describen los principales métodos utilizados en el trabajo.

[1]Instituto de Estandarización de los EE.UU., encargado de validar y establecer los métodos que representan el estado del arte en el RAL mediante las evaluaciones SRE (Speaker Recognition Evaluation).

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 3, Julio-Septiembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 137-147
http://rcci.uci.cu

Pre-procesamiento de la señal de audio: Filtro de Wiener

El Filtro de Wiener es un filtro propuesto por Norbert Wiener en la década de 1940 y publicado en 1949. Su propósito es reducir el ruido aditivo presente en la señal observada utilizando métodos estadísticos, de tal modo que la señal estimada a la salida del filtro se aproxime lo más posible a una señal deseada sin ruido. El filtro produce un estimado de la señal deseada aplicando un filtro lineal e invariante en el tiempo de la señal ruidosa observada, asumiendo conocidos el espectro de la señal y del ruido aditivo, minimizando el error medio cuadrático entre la señal estimada y la señal deseada. El filtro de Wiener se caracteriza por:

- Se asume que la señal observada contiene ruido aditivo, que la señal y el ruido son procesos estocásticos lineales y estacionarios, que se conocen sus características espectrales, su auto-correlación y su cros-correlación.
- El filtro debe ser físicamente realizable y causal.
- El criterio de comportamiento es el MMSE: error medio-cuadrático mínimo.

Análisis del comportamiento del Filtro de Wiener

El análisis del comportamiento del Filtro de Wiener aplicado sobre diversos tipos de señales, se realizó basándonos en la SNR[2] de las señales filtradas y sin filtrar, y se comparó con un filtro de Sustracción Espectral[3]. El Filtro de Wiener se implementó apoyándonos en la propuesta hecha en los estándar ETSI ("European Telecommunications Standards Institute") 202-212 v1.1.2 del 2005 y 202-050 v1.1.5 del 2007, donde especifican los algoritmos para la extracción de características de la voz "ETSI advanced front-endz su transmisión, como parte de un sistema distribuido de reconocimiento de voz.

La tabla 1 muestra los parámetros característicos de las señales analizadas, las SNR de la señal original y las SNR posterior al filtrado:

- Señal: nombre de la señal utilizada, las cuales presentan diferentes tipos de ruido ambiental, tomadas de bases internacionales.
- Duración: duración aproximada de las señales, en segundos.

[2]Para medir las SNR de las señales de voz se utilizó la medida propuesta por la compañía rusa STC en su herramienta (Test Sound Processing).
[3]Propuesto e implementado por la compañía rusa STC.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 3, Julio-Septiembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 137-147
http://rcci.uci.cu

- Fm: frecuencia de muestreo de las señales, en Hz.
- # bits: número de bits por muestra
- SNR: relación señal-ruido de la señal original.
- SNR-SE: SNR de la señal filtrada con el filtro de Sustracción Espectral.
- SNR-STD: SNR de la señal estandarizada a 8000 Hz,16 bits.
- SNR-SRD+Wiener: SNR de la señal estandarizada a 8000 Hz,16 bits y posteriormente filtrada con el filtro de Wiener.

Tabla 1. Resultados de la aplicación del Filtro de Wiener sobre señales con diferentes tipos de ruido ambiental tomadas de bases de datos internacionales y brindando como medida la relación señal ruido SNR.

Señal	Duración	Fm	# bits	SNR	SNR-SE	SNR-STD	SNR-STD+Wiener
Airport8k	180	8000	8	2,3	20,7	3,2	17,4
Bable8k	240	8000	8	1,2	16,6	1,2	11,8
Car8k	22	8000	8	-1,8	5,4	-1,8	10,5
Exhibition8k	19	8000	8	0,7	17,5	0,7	14,5
Exhibition16k	19	16000	16	0,6	17,5	0,7	14,5
Noises8k	80	8000	8	8	27,9	8,0	12,2
Restaurant8k	285	8000	8	3,5	21,1	3,5	17,1
Restaurant16k	285	16000	16	3,4	21,1	3,5	17.0
Street8k	60	8000	8	6,4	19,3	6,4	21,1
Street16k	60	16000	16	4,7	19,3	6,4	21,1
Train8k	180	8000	8	1,9	14,2	1,9	15,1

A partir del análisis de la tabla 1 podemos decir que en casi todas las señales, el filtro de sustracción espectral obtiene una SNR más elevada que el filtro Wiener, pero incorporando el ruido musical, el cual aporta ininteligibilidad al segmento de voz. El filtro Wiener no logra alcanzar las mismas SNR pero las señales quedan más inteligibles. Incluso en el caso de las señales con diferentes tipos de ruido ambiental, que son muy ruidosas, el filtro de Wiener logra elevar la SNR entre 10 y 20 dB. Por otra parte se puede observar que la estandarización a 8000 Hz y 16 bits de todas las señales, no aporta al mejoramiento de la SNR.

Todo lo anterior, unido a su utilización en las competencias NIST-SRE 2012, nos hace pensar que el filtro de Wiener es más adecuado que el filtro de sustracción espectral, para el pre-procesamiento de las muestras celulares y telefónicas en el RAL.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 3, Julio-Septiembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 137-147
http://rcci.uci.cu

Representación del locutor y compensación de variabilidad de sesión

Los sistemas de RAL actuales se basan en la representación i-vector para representar la información discriminatoria del locutor. La representación i-vector puede definirse mediante una distribución a posteriori de las variables ocultas, condicionadas a las estadísticas de 0 y 1 orden de Baum-Welch extraídas del segmento de voz. El i-vector se obtiene a partir de un único espacio de variabilidad denominado Espacio de Variabilidad Total (T) (Dehak et al., 2011), que contiene simultáneamente las variabilidades del locutor y de la sesión. Esta representación del locutor se formula por

$$M = m + Tw, \tag{1}$$

donde m es un supervector obtenido mediante la concatenación de los vectores de media del Modelo Universal de Fondo (UBM), que contiene la información independiente del locutor y de la sesión, T es una matriz rectangular de bajo rango y w es un vector intermedio que sigue una distribución normal $\mathcal{N}(0, I)$ y representa la información discriminatoria del locutor, denominado i-vector. En la ecuación 1, se asume que el vector M mantiene una distribución normal con m and TT' como media y covarianza respectivamente.

La variabilidad de sesión es conocida por ser un factor importante en la degradación de la eficacia del RAL. La reducción de esta variabilidad representa una parte obligatoria de los sistemas actuales de RAL. Algunos métodos de compensación o reducción de variabilidad se han venido aplicando con el objtivo de aumentar la eficacia del reconocimiento, dentro de este grupo el más utilizado es el Análisis Discriminante Linear (LDA) (Rao, 1948). El método LDA es una técnica de reducción de dimensionalidad que actualmente es aplicada en la rama del reconocimiento de locutores sobre el espacio de los i-vectores, con el proposito de compensar la variabilidad de sesión (Dehak et al., 2011). El objetivo principal de aplicar el LDA, es poder maximizar la dispersión entre las clases (S_b) y simultaneamente minimizar la dispersión dentro de la clase (S_w), partiendo de una población de locutores.

$$S_b = \sum_{l=1}^{L} (x_l - \bar{x})(x_l - \bar{x})', \tag{2}$$

donde L es la cantidad de locutores de la población, x_l es la media de los i-vectores por cada locutor y $\bar{x}$ es el vector de media global dado una población de locutores, y

$$S_w = \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i=1}^{n_l} (x_i^l - x_l)(x_i^l - x_l)', \tag{3}$$

donde n_l es la cantidad de i-vectores del locutor l y x_i^l es el i-th i-vector del locutor l.

La matriz de proyección A, es un subconjunto de vectores propios J asociados a los mayores valores propios, los cuales son obtenidos mediante la optimización del criterio de Fisher:

$$J(v) = \frac{v' S_b v}{v' S_w v}, \tag{4}$$

donde v es la dirección del espacio dado.

La medida de similitud del coseno entre dos i-vectores w_1 y w_2, cuando se compensa la variabilidad de sesión con el método LDA, se define mediate:

$$CSM_{(w_1,w_2)} = \frac{(A'w_1)'(A'w_2)}{||A'w_1||||A'w_2||}. \tag{5}$$

Diseño de los experimentos

Para evaluar el comportamiento del Filtro de Wiener aplicado al RAL nos apoyamos en la representación i-vector usando como medida de similitud la distancia del coseno. Y se propuso dos configuraciones de evaluación apoyados en las sesiones telefónicas masculinas de la base de voces NIST SRE-08 (Gonzalez-Rodriguez, 2014):

1. ruido-ruido: se realiza una evaluación de RAL sobre condiciones ruidosas en los segmentos de voz del cliente y los segmentos de voz de identidad desconocida.
2. limpio-ruido: se realiza una evaluación utilizando solamente muestras de voces ruidosas en los segmentos de voz de identidad desconocida.

Los detalles en las configuraciones de los experimentos de reconocimiento de locutores se describen a continuación en las siguientes secciones.

Bases de Voces y extracción de rasgos acústicos

El entrenamiento del modelo UBM, matriz T y matriz de compensación LDA se realizó utilizando un conjunto de segmentos de voces masculinas telefónicas, extraidas de las bases de voces NIST SRE-04 y SRE-05 (Gonzalez-Rodriguez, 2014). La evaluaciones de verificación de locutores se realizaron sobre un conjunto de segmentos de voces extraidos de las sesiones telefónicas masculinas, `short2` y `short3`; de la base NIST SRE-08. Para un total de 470 clientes y 670 segmentos de voces a verificar, lo que representa un total de 6615 verificaciones.

Revista Cubana de Ciencias Informáticas
Vol. 12, No. 3, Julio-Septiembre, 2018
ISSN: 2227-1899 | RNPS: 2301
Pág. 137-147
http://rcci.uci.cu

Las señales corruptas por ruido se obtuvieron simulando el conjunto de evaluación a partir de diferentes ruidos como: ruidos propios de restaurantes, de calles, ruidos en aereopuertos, automóviles, cafeterias y ruido comunes en una exhibición. Estas muestras se simularon utilizando la herramienta pública FaNT[4] y con diferentes relaciones señal a ruido (SNR) entre 2 y 20 db.

Para representar el espectro del habla a corto término se utilizaron los Coeficientes Cepstrales en Frecuencia Lineal (LFCC) (Scheffer et al., 2013), con una dimensión de 50 coeficientes por cada 10 mseg de voz. Para eliminar el silencio existente en la señales de voz se aplicó un algoritmo de detección de la actividad de voz (VAD) (Sohn et al., 1999). Finalmente los rasgos acústicos LFCC se normalizan en función de su media y varianza (CMVN).

Configuración del UBM, T y LDA

Se entrenó un modelo UBM de 512 componentes gaussianas basado en la covarianza diagonal y una matriz T de rango 400, utilizando 3911 segmentos de voz de 262 locutores masculinos. Por cada segmento de voz se extrae un i-vector de 400 dimensiones a partir de las estadísticas de 0 y 1^{er} orden de Baum-Welch. La matriz de compensación de la variabilidad de sesión se entreno utilizando el algoritmo LDA con una reducción de dimensión hasta 250.

Resultados y discusión

En la tabla 2 se muestran los resultados de las evaluaciones realizadas para analizar el comportamiento del RAL sobre escenarios ruidosos cuando se le aplica o no el Filtro de Wiener. Nos apoyamos para medir la eficacia del reconocimiento de locutores en el error equiprovable (% EER) y el mínimo de la función de costo (minDCF) (Gonzalez-Rodriguez, 2014).

Tabla 2. Resultados de la evaluación en la verificación de locutores bajo diferentes condiciones de ruidos y SNR en función del % de error equiprobable (EER) y el mínimo de la función de costo (minDCF)

	Sin Filtro de Wiener		Con Filtro de Wiener	
Condiciones	**% EER**	**minDCF*100**	**% EER**	**minDCF*100**
`limpio-limpio`	4,55	2,34	—	—
`ruido-ruido`	14,6	6,31	13,9	6,10
`limpio-ruido`	9,11	4,12	7,97	3,88

[4]http://dnt.kr.hsnr.de/download.html

Como se observa en la tabla 2, los métodos de reconocimiento de locutores sufren un drástico aumento del error cuando es aplicado sobre escenarios ruidosos, decreciendo su eficacia en un 68,8 % y 50,0 % respectivamente para ambas condiciones de evaluación. Por otra parte, podemos observar que el sistema RAL obtiene resultados superiores en la condición de evaluación `limpio-ruido` respecto a la `ruido-ruido`. Estos resultados esta dado a que la muestra de voz limpia presenta mayor contenido de información discriminatoria al locutor que la muestra de voz ruidosa.

Por otra parte la aplicación del Filtro de Wiener para mejorar la calidad de los segmentos de voz en los sistemas de RAL, robustece los métodos de reconocimiento alcanzando una mejora relativa del EER de un 4,94 % y 12,5 % respectivamente.

Conclusiones y Trabajo Futuro

Este trabajo realiza un análisis del comportamiento del filtro de Sustracción Espectral y el Filtro de Wiener en la reducción de ruido sobre señales de voz, con el objetivo de aplicar el más robusto entre ellos en la fase de pre-procesamiento de las señales de audio del sistema de RAL, en aras de aumentar la eficacia en escenarios ruidosos. Este análisis obtuvo como resultado que el Filtro de Wiener presenta un grupo de ventajas en relación a la Sustracción Espectral, principalmente, el filtrado de Wiener no incorpora ruido musical a las muestras ruidosas que procesan.

Por otra parte, la aplicación del Filtro de Wiener en la fase de pre-procesamiento de las señales del sistemas de RAL, aumenta la eficacia de los métodos de reconocimiento bajo condiciones adversas de ruido en un 4,94 % y 12,5 % de mejora relativa.

No obstante, si se tiene en cuenta que, los rasgos robustos para enfrentar el ruido reducen la eficacia de los métodos de reconocimiento de locutores en presencia de señales limpias, nos proponemos como trabajo futuro realizar un análisis del comportamiento del Filtro de Wiener ante segmentos de voz que no están afectados por ruido.

Referencias

Anshu Agarwal and Yan Ming Cheng. Two-stage mel-warped wiener filter for robust speech recognition. In *Proc. ASRU*, volume 99, pages 67–70, 1999.

Hynek Boril, Frantisek Grézl, and John HL Hansen. Front-end compensation methods for lvcsr under lombard effect. In *INTERSPEECH*, pages 1257–1260, 2011.

Gillian M Davis. *Noise reduction in speech applications*, volume 7. CRC Press, 2002.

N. Dehak, P. Kenny, R. Dehak, Dumouchel P., and Ouellet P. Front-end factor analysis for speaker verification. volume 19, pages 788–798, 2011.

Luciana Ferrer, Mitchell McLaren, Nicolas Scheffer, Yun Lei, Martin Graciarena, and Vikramjit Mitra. A noise-robust system for nist 2012 speaker recognition evaluation. Technical report, SRI INTERNATIONAL MENLO PARK CA SPEECH TECHNOLOGY AND RESEARCH LAB, 2013.

Joaquin Gonzalez-Rodriguez. Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014). *Loquens*, 1(1), 2014.

Craig S Greenberg, Vincent M Stanford, Alvin F Martin, Meghana Yadagiri, George R Doddington, John J Godfrey, and Jaime Hernandez-Cordero. The 2012 nist speaker recognition evaluation. In *INTERSPEECH*, pages 1971–1975, 2013.

Chanwoo Kim and Richard M Stern. Power-normalized cepstral coefficients (pncc) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(7):1315–1329, 2016.

Miranti Indar Mandasari, Mitchell McLaren, and David A van Leeuwen. The effect of noise on modern automatic speaker recognition systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4249–4252. IEEE, 2012.

Mitchell McLaren, Yun Lei, Nicolas Scheffer, and Luciana Ferrer. Application of convolutional neural networks to speaker recognition in noisy conditions. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Ji Ming, Timothy J Hazen, James R Glass, and Douglas A Reynolds. Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1711–1723, 2007.

Vikramjit Mitra, Horacio Franco, Martin Graciarena, and Arindam Mandal. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4117–4120. IEEE, 2012.

Timur Pekhovsky, Sergey Novoselov, Aleksei Sholohov, and Oleg Kudashev. On autoencoders in the i-vector space for speaker recognition. In *Proc. Odyssey*, 2016.

Oldrich Plchot, Lukas Burget, Hagai Aronowitz, and Pavel Matëjka. Audio enhancing with dnn autoencoder for speaker recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5090–5094, 2016.

Padmanabhan Rajan, Tomi Kinnunen, and Ville Hautamäki. Effect of multicondition training on i-vector plda configurations for speaker recognition. In *Interspeech*, pages 3694–3697, 2013.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

Fred Richardson, Brian Nemsick, and Douglas Reynolds. Channel compensation for speaker recognition using map adapted plda and denoising dnns. In *Proc. Speaker Lang. Recognit. Workshop*, pages 225–230, 2016.

Seyed Omid Sadjadi and John HL Hansen. Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5448–5451. IEEE, 2011.

Rahim Saedi, Kong Aik Lee, Tomi Kinnunen, Tawfik Hasan, Benoit Fauve, Pierre-Michel Bousquet, Elie Khoury, Pablo Luis Sordo Martinez, Jia Min Karen Kua, Changhuai You, et al. I4u submission to nist sre 2012: A large-scale collaborative effort for noise-robust speaker verification. In *Interspeech*, number EPFL-CONF-192763, 2013.

Nicolas Scheffer, Luciana Ferrer, Aaron Lawson, Yun Lei, and Mitchell McLaren. Recent developments in voice biometrics: Robustness and high accuracy. In *Technologies for Homeland Security (HST), 2013 IEEE International Conference on*, pages 447–452. IEEE, 2013.

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3, 1999.