UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA
Tecnología & Desarrollo
Vol. 14, Nº 1, Enero - Diciembre 2016, Pág.097-100

Atributos PNCC para reconocimiento robusto de locutor independiente del texto

PNCC attributes for robust recognition of independent speaker from the text

Harry Anacleto Silva[1]

RESUMEN

El reconocimiento automático de locutores ha sido sujeto de intensa investigación durante toda la década pasada. Sin embargo las características, del estado de arte de los algoritmos son drásticamente degradados en presencia de ruido. Este artículo se centra en la aplicación de una nueva técnica llamada Power-Normalized Cepstral Coefficients (PNCC) para el reconocimiento de locutor independiente del texto. El objetivo de este estudio es evaluar las características de esta técnica en comparación con la técnica convencional Mel Frequency Cepstral Coefficients (MFCC) y la técnica Gammatone Frequency Cepstral Coefficients (GFCC).

Palabras clave: Reconocimiento de Locutor, MFCC, GFCC, PNCC, Robustez frente al ruido.

ABSTRACT

Automatic speaker recognition has been a subject of intense research over the past decade. However, its performance under state-of-art algorithms is drastically degraded in presence of noise. This paper focuses on the application of the novel technique Power-Normalized Cepstral Coefficients (PNCC) to the text-independent speaker recognition task. Our aim of this study is to evaluate the performance of this feature in comparison with the conventional Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC) methods.

.

Keywords: Speaker identification, MFCC; GFCC; PNCC, Noise robustness, Speaker identification, MFCC, GFCC, PNCC, noise robustness.

Recepción: 15-07-2016 / **Aceptación:** 08-08-2016

[1] Universidad César Vallejo - Escuela de Ingeniería de Sistemas. E-mail: hanacleto@ucv.edu.pe

1. INTRODUCCIÓN

El objetivo de la identificación de locutores es usar la señal de voz para identificar a quien de los locutores registrados pertenece dicha voz. Sin embargo, las características del reconocimiento son severamente degradadas en ambientes con ruido, especialmente en condiciones de SNR muy bajas. Estrategias robustas pueden ser desarrolladas en diferentes etapas del procesamiento tales como: en el pre-procesamiento, en la extracción de las características, en el sistema de clasificación o en la elección de la medida de similaridad.
Este articulo se centra en la aplicación de la nueva técnica de atributos PNCC[1]para la tarea de reconocimiento de locutor independiente del texto. Para propósitos de comparación fueron considerados los atributos GFCC (Gammatone Frequency Cepstral Coefficients), que ha proporcionado buenos resultados para el reconocimiento robusto de locutor[2].
El resto de esta investigación se organiza de la siguiente manera: sección 2 describe brevemente los conceptos relacionados con MFCC, GFCC y su análisis de características, sección 3 describe la nueva técnica PNCC con los resultados de su validación experimental, finalmente, las conclusiones son dadas en la sección 4.

2. DESCRIPCIÓN DE CARACTERÍSTICAS Y ANÁLISIS DE FUNCIONAMIENTO

En esta sección se describe brevemente los Atributos de voz utilizados en el presente artículo y su análisis de funcionamiento.

A. Descripción de los Atributos

A.1. Mel Frequency Cepstral Coefficients (MFCC)

El proceso de la extracción de los atributos MFCC es el siguiente[2]:

1. Pre-énfasis en la señal de entrada.

2. Aplicar la transformada de Fourier de tiempo reducido para obtener el espectro de magnitud.

3. Obtener el espectro Mel del espectro de magnitud usando 26 filtros triangulares con superposición de ventanas donde el centro de frecuencias están igualmente distribuidos sobre la escala mel.

4. Obtener el logaritmo de la potencia del espectro (i.e. cuadrado del. espectro Mel).

5. Aplicar la transformada discreta del coseno (DCT) del logaritmo del espectro de potencia Mel para obtener los coeficientes MFCC.

A.2. Gammatone Frequency Cepstral Coefficients (GFCC)

El proceso de la extracción de los atributos GFCC sigue los mismos pasos que [3] con cambios en la etapa 3:

1. Pasar la señal por un banco de de filtros Gammatone de 64 canales.

2. En cada canal: i) rectificar totalmente la respuesta del filtro. ii) decimar a 100 Hz la señal y tomar el valor absoluto para crear una representación (T-F) que es una variante del cocleagrama.

3. Obtener el logaritmo de la representación T-F.

4. Aplicar DCT para obtener los coeficientes GFCC.

B. Análisis de funcionamiento

A fin de evaluar las características de esos tres métodos para reconocimiento de locutor, fue utilizada la base de datos TIMIT, que contiene 630 locutores de los 8 más importantes dialectos del Ingles Americano, donde el 70% son hombres (438) y los 30% restantes son mujeres (192), cada locutor contribuye con 10 frases de 3 segundos como promedio de duración[4].
En esta investigación, 330 locutores fueron escogidos aleatoriamente, y por cada locutor 8 frases fueron usadas para la etapa de entrenamiento y las 2 frases restantes para la etapa de pruebas[2].
Fueron definidos 3 escenarios para producir los resultados. En cada unos de estos escenarios la selección del locutor así como las frases usadas para la etapa de entrenamiento y de pruebas fueron cambiando de la siguiente manera:

- **Escenario 1** - Para cada una de las 10 frases habladas por cada locutor, 8 frases fueron escogidas para entrenamiento y las restantes 2 frases para la etapa de pruebas.
- **Escenario 2** - Para cada una de las 10 frases habladas por cada locutor, 2 sentencias fueron escogidas para la etapa de pruebas, las cuales son diferentes que las frases obtenidas en el escenario 1 y las restantes 8 frases fueron usadas para entrenamiento.
- **Escenario 3** - Para cada una de las 10 frases habladas por cada locutor, fueron escogidas 2 frases para la etapa de pruebas, diferentes que las frases escogidas en el escenario 1 y 2, y las restantes 8 frases fueron escogidas para entrenamiento.

Los datos limpios de la etapa de pruebas, fueron mezclados con 3 tipos de ruido: factory, pink y white de la base de datos NOISEX-92 con una relación señal ruido (SNR) de 0 y 5 dB.

3. EL USO DEL ATRIBUTO POWER-NORMALIZED CEPSTRAL COEFFICIENTS (PNCC)

A continuación es presentada la nueva técnica llamada PNCC para el reconocimiento automático de locutor independiente del texto, la cual ha demostrado buenos resultados para el reconocimiento de voz. Esta técnica emplea los siguientes pasos para el proceso de extracción de características[1]:

1. Pre-énfasis en la señal de entrada.
2. Aplicar la transformada de Fourier de tiempo reducido para obtener el espectro de magnitud.
3. El espectro es divido en 40 bandas usando un banco de filtros Gammatone, donde el centro de frecuencias están igualmente distribuidos en la Equivalent Rectangular Bandwidth (ERB) escala.
4. El ruido en cada banda es estimada y removida.
5. La energía en cada banda es calculada con el exponente 1/15.
6. Aplicamos DCT para obtener los atributos PNCC.

Para los atributos MFCC y PNCC, un filtro de pre énfasis de la forma $H(z) = 1 - 0.97z^{-1}$ es aplicado. Los atributos de la técnica MFCC fueron obtenidos cada 16 ms con 4 ms entre cuadros y empleado 26 filtros triangulares con sus centro de frecuencias uniformemente distribuida en la escala Mel entre 50 Hz y 8000 Hz. Los atributos of PNCC fueron extraídos cada 25.6 ms con 10 ms entre cuadros. Para los atributos PNCC y GFCC fueron empleados 40 y 64 filtros Gammatone, uniformemente distribuido en la escala ERB (Equivalent Rectangular Bandwidth) entre 200 Hz y 8000 Hz.

Es utilizada máscara binaria ideal (IBM) para la detección de actividad de voz[5], y por eso, cada uno de los atributos propuestos deben tener una representación T-F (tiempo-frecuencia). De este modo, la representación T-F de MFCC y PNCC provienen de la potencia del espectro, donde cada elemento representa la energía de la unidad T-F correspondiente. Los cuadros con al menos una confiable unidad T-F (cuando la energía no es dominada por el ruido), es etiquetado como 1 IBM, y se seleccionan para el reconocimiento.

Veintidós coeficientes dimensionales PNCC, MFCC y GFCC, con el coeficiente 0 removido, son usados para este estudio. Los locutores son modelados usando el Modelo de Mistura Gaussiana (GMM) con 32 misturas[6].

Las Tablas 1 a 3 presentan la comparación del desempeño entre las técnicas PNCC, MFCC y GFCC para diferentes tipo de ruido y diferentes niveles de SNR. Estas tablas muestran que el desempeño de PNCC es mucho más alto que GFCC y MFCC en cada uno de los tres escenarios con diferentes tipos de ruido y SNR. Puede ser observado que el desempeño de MFCC es muy bajo en el caso de ruido white.

Tabla 1. Desempeño de GFCC, MFCC y PNCC en terminos de porcentajes de reconocimiento de locutor (%)

Ruído	SNR	Escenario 1		
	(dB)	GFCC	MFCC	PNCC
Factory	0	1.82	4.24	31.21
	5	26.06	25.45	76.67
Pink	0	1.52	1.82	21.52
	5	4.85	4.85	53.94
White	0	0.91	0.30	6.97
	5	3.94	1.82	26.06

Fuente: Elaboración propia.

Tabla 2. Desempeño de GFCC, MFCC y PNCC en términos de porcentajes de reconocimiento de locutor (%).

Ruído	SNR	Escenario 2		
	(dB)	GFCC	MFCC	PNCC
Factory	0	2.12	2.73	36.97
	5	25.45	26.97	78.18
Pink	0	0.30	0.67	25.76
	5	8.48	4.04	61.21
White	0	1.52	0.91	10.61
	5	4.85	1.82	33.03

Fuente: Elaboración propia.

Tabla 3. Desempeño de GFCC, MFCC y PNCC En términos de porcentajes de reconocimiento de locutor (%).

Ruído	SNR	Escenario 3		
	(dB)	GFCC	MFCC	PNCC
Factory	0	3.03	9.09	34.24
	5	32.73	30.91	73.64
Pink	0	1.21	1.52	24.55
	5	8.79	6.97	54.55
White	0	0.91	0.61	10.30
	5	3.33	2.73	28.48

Fuente: Elaboración propia.

4. CONCLUSIONES

Este artículo propone el uso del atributo PNCC para el reconocimiento robusto de locutor. El uso del atributo PNCC, para el reconocimiento del locutor es una técnica eficaz que permite obtener mejores resultados en comparación con los métodos tradicionales que se vienen aplicando. La superioridad de PNCC sobre MFCC y GFCC ha sido confirmada con las pruebas que se han aplicado, demostrando su efectividad incluso para voz distorcionada con ruido white, pink y factory para SNR bajas. Siendo que el MFCC tiene un desenvolvimiento pobre para los ruidos pink y white.

5. REFERENCIAS BIBLIOGRÁFICAS

[1] Kim C ,Stern R, Feature Extraction For Robust Speech Recognition Based On Maximizing The Sharpness Of The Power Distribution And On Power Flooring. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. March 2010: 4574–4577.

[2] Zhao X , Wang D ,Analyzing Noise Robustness Of Mfcc And Gfcc Features In Speaker Identification. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. May 2013: 7204–7208.

[3] Zhao X, Shao Y , Wang DL. CASA - Based Robust Speaker Identification. IEEE Trans. Audio, Speech and Language Processing. 2012; 20(5) :1608-1616.

[4] Garofolo JS, Lamel LF , Fisher WM, Fiscus JG, Pallett DS , Dahlgren NL.The DARPA - TIMIT Acoustic-Phonetic Continuous Speech Corpus. [CD-ROM] Philadelphia: Linguistic Data Consortium.1993

[5] Wang DL , Brown GJ, Computational Auditory Scene Analysis: Principles, algorithms, and applications. Hoboken. New Jersey: Wiley-IEEE; 2006.

[6] Reynolds D ,Rose R, Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models., IEEE Trans. on Speech and Audio Processing.1995; 3: 72-83.