Comparación de Algoritmos de Aprendizaje para Identificación del Usuario a través de la Voz

Esaú Villatoro, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez
Instituto Nacional de Astrofísica, Óptica y Electrónica
{villatoroe, villasen, mmontesg}@inaoep.mx

RESUMEN
En este trabajo presentamos una comparación entre cuatro algoritmos de aprendizaje automático para identificación del hablante. El estudio hace hincapié en la simplificación de la caracterización de la señal de voz al no usar reconocimiento fonético. Los resultados hasta ahora alcanzados nos brindan elementos para preferir el algoritmo de Máquinas de Vectores de Soporte (SVM).

Palabras Claves
Personalización de la interfaz, identificación del hablante, máquinas de vectores de soporte.

INTRODUCCIÓN
Una interfaz dinámica que cambie su apariencia o que proponga los elementos más comúnmente utilizados por un usuario específico son algunos de los elementos que se buscan al personalizar una interfaz. Por supuesto, uno de los primeros pasos en esta tarea es identificar al usuario. Este trabajo está orientado a identificar al usuario a través de su voz. De manera primordial el habla conlleva un mensaje a través del idioma. Pero la señal de voz también nos permite determinar el lenguaje hablado, la emoción, el género, en ocasiones la edad, así como la identidad del hablante. La tarea de *identificación del hablante* es una tarea que ha sido ampliamente estudiada, y además es desempeñada con altos porcentajes de certeza[1]. En nuestro caso deseamos identificar al usuario sin depender de lo que se diga, es decir, buscamos un sistema de identificación *independiente* del texto. Por otro lado, uno de los retos del presente trabajo es realizar la tarea de identificación sin depender de un reconocedor fonético, al trabajar directamente sobre los coeficientes extraídos de la señal de voz. A continuación se describen los experimentos realizados así como los resultados alcanzados.

EXPERIMENTOS
Para las fases de entrenamiento y evaluación se usó un corpus para el español mexicano conformado por grabaciones de 50 personas. Por cada persona se tienen 50 frases a 44kHz sin ruido, de 3 seg. La duración total del corpus es de aproximadamente 250 minutos. Cada grabación fue segmentada en ventanas sin solapamiento de: 30ms, 50ms y 80ms. De cada segmento se extrajeron sus coeficientes utilizando LPC (*Linear Predictive Coefficients*) y MFCC (*Mel Frequency Cepstra Coefficients*). Los algoritmos de aprendizaje automático usados en esta comparación fueron: redes neuronales (ANN), vecinos más cercanos (k-nn), regresión lineal localmente ponderada (LWR) y máquinas de vectores de soporte (SVM). Las configuraciones específicas de cada algoritmo fueron: *k-nn* con k igual 50 e igual a 60; para LWR k=2000; para ANN tipo *feedforward* con 200:125:50 (200 nodos de entrada, 125 nodos en la capa oculta, 50 nodos para las clases), entrenadas por 1600 épocas con una tasa de aprendizaje de 0.005; y para SVM no se hizo ningún tipo de re-escalamiento a los datos y se utilizó un kernel de tipo polinomial. Para evaluar cada uno de los métodos de aprendizaje se utilizó la técnica de validación cruzada con 10 pliegues (*10 fold cross-validation*).

LEAVE BLANK THE LAST 2.5 cm (1") OF THE LEFT COLUMN ON THE FIRST PAGE FOR THE COPYRIGHT NOTICE.

RESULTADOS
La siguiente tabla muestra la precisión alcanzada con cada algoritmo para cuando se caracterizó la señal con MFCC y LPC.

	Algoritmo	***30ms***	***50ms***	***80ms***
MFCC	knn (50)	27.09%	28.61%	29.09%
	knn (60)	26.20%	28.41%	28.53%
	LWR	55.46%	55.98%	56.35%
	SVM	63.50%	63.66%	63.82%
	ANN	48.83%	50.04%	47.91%
LPC	knn (50)	9.47%	9.83%	10.19%
	knn (60)	9.67%	8.55%	10.19%
	LWR	30.91%	34.64%	42.99%
	SVM	31.52%	30.63%	33.92%
	ANN	20.64%	21.47%	24.48%

Como puede observarse el mejor comportamiento se consigue al caracterizar la señal de voz usando MFCC y bajo este caso el mejor algoritmo es SVM. Este método se comporta mejor dada la complejidad de los atributos y su difícil representación en un espacio lineal. También se distinguen mejores resultados al utilizar segmentos de 80 ms.

AGRADECIMIENTOS
El presente trabajo fue realizado con el apoyo parcial del Laboratorio Franco-Mexicano de Informática.

REFERENCIAS
1. Reynolds, D.A.: An Overview of Automatic Speaker Recognition Technology. *ICASSP*, 2002.