Identificación Automática del Hablante mediante Redes Neuronales

Conference Paper · June 1999

CITATION
1

READS
63

2 authors:

Humberto Maximiliano Torres
National Scientific and Technical Research Council
42 PUBLICATIONS **141** CITATIONS
SEE PROFILE

Hugo Leonardo Rufiner
Universidad Nacional del Litoral
115 PUBLICATIONS **835** CITATIONS
SEE PROFILE

Identificación Automática del Hablante mediante Redes Neuronales

Torres Humberto M., Rufiner Hugo L.

Facultad de Ingeniería – Bioingeniería
Universidad Nacional de Entre Ríos

Resumen

El presente trabajo consiste en la utilización de Coeficientes Delta Cepstra en escala de Mel para alimentar un sistema de identificación automática del hablante basado en redes neuronales. Se ensayan diferentes alternativas para el clasificador basado en redes neuronales, lográndose muy buenos desempeños para conjuntos cerrados de hablantes en forma independiente del texto.

Identificación del Hablante • Análisis de voz • Redes neuronales

1. Introducción

La señal de la voz tiene distintos niveles de información. Primeramente, lleva las palabras o mensajes, pero en un nivel secundario, la señal también lleva consigo información acerca de la identidad del hablante. Mientras que el área del reconocimiento automático del habla es relativo a la extracción del mensaje lingüístico en una oración, el área del reconocimiento del hablante es concerniente con la extracción de la identidad de la persona [1].

El reconocimiento automático del habla es un campo multidisciplinar con especial vinculación a la informática y, dentro de ella y de forma especial, al reconocimiento de formas y la inteligencia artificial [2].

La tarea de identificación automática del hablante es la determinación por parte de una máquina de la identidad del hablante. Para que una persona pueda reconocer una voz, esta debe ser familiar, algo similar ocurre con las máquinas. El proceso de aprender las características particulares de un hablante se conoce como entrenamiento, y utiliza una colección de datos (oraciones) de la persona a ser identificada. La segunda etapa o componente de la identificación del hablante es el de prueba; esta tarea consiste en comparar oraciones desconocidas con los datos de entrenamiento y hacer la identificación propiamente dicha.

De acuerdo a las frases utilizadas en el reconocimiento del hablante, este se divide generalmente en dos clases: (1) dependiente del texto y (2) independiente del texto. En el modo dependiente del texto el hablante provee las mismas oraciones, tanto para entrenamiento como para prueba. Una aproximación clásica de identificación del hablante dependiente del texto es por reconocimiento de patrones, donde usualmente son aplicados métodos de programación dinámica [3] para reconocer temporalmente las oraciones de prueba y entrenamiento. Además, a causa de que la misma secuencia de eventos fonéticos son pronunciados, las características dinámicas del espectro son frecuentemente incluidas en las características extraídas para proveer reconocimiento. Por otra parte, la identificación del hablante independiente del texto no impone que el texto de entrenamiento y reconocimiento sea el mismo. Una opción intermedia consiste en una aproximación texto-dirigido, la cual es ahora comúnmente adoptada. Aquí el usuario es consultado por alguna frase u oración la cual es aleatoriamente seleccionada por el sistema y no puede ser predicha con anterioridad [4].

Otra división posible de la identificación del hablante, es en cuanto si se trata de un problema de conjunto cerrado o un problema de conjunto abierto.

En el problema de conjunto cerrado se trata de identificar un hablante de entre un grupo de N posibles hablantes. Naturalmente, el valor elevado de N, aumenta la dificultad del reconocimiento. En este caso, el hablante que tenga características más parecidas a la oración de prueba, será identificado con esta.

En el problema de conjunto abierto, se trata de definir si una oración pertenece o no a un hablante de un grupo de N hablantes. En este caso no se trata de definir si es un hablante en particular, sino que es suficiente con que pertenezca a un hablante del grupo, se trata de una decisión binaria (pertenece o no pertenece al grupo). Este problema no es necesariamente más fácil que el anterior. La verificación del hablante es un caso especial del problema de conjunto abierto, y se trata de decidir si la persona es la que dice ser.

El problema de identificación del hablante puede ser dividido en dos componentes: análisis de la voz (o extracción de características) y clasificación.

Las RNA son excelentes sistemas de clasificación y se especializan en trabajar con datos ruidosos, incompletos, solapados, etc. El problema de identificación del hablante es una tarea de clasificación de datos que tiene todas estas características, haciendo a las RNA una alternativa atractiva a la aproximación descripta.

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. M. Torres & H. L. Rufiner; "Identificación Automática del Hablante mediante Redes con Retardos"
Anales del XII Congreso Argentino de Bioingeniería. Vol. 1, pp. 319--322, Jun, 1999.

2. Material y Métodos

2.1 Acerca de los datos

Las señales de voz para los experimentos de identificación fueron obtenidas del corpus de voz continua TIMIT **[5]**. Esta base de datos ha sido confeccionada en forma conjunta por Texas Instruments (TI) y el Massachusetts Institute of Technology (MIT). Es una de las bases multi-hablante más empleadas en el ámbito del Reconocimiento Automático del Habla (RAH) del discurso continuo por ser la más grande, completa y mejor documentada de su tipo. Esta base o corpus posee una gran cantidad de fonemas en diversos ambientes y pronunciados por más de 600 hablantes diferentes. El corpus TIMIT incluye la señal de voz correspondiente a cada oración hablada, así como también las transcripciones ortográficas, fonéticas y de palabras alineadas temporalmente.

2.2 Organización de los datos

Se debe observar que, para nuestro trabajo, la separación original de las oraciones de TIMIT en entrenamiento y prueba no es válida, ya que los hablantes de uno y otra son distintos. Ante esto, se decidió tomar tres de cada cuatro oraciones, por hablante, para entrenar y la restante para prueba. Esta división disminuye la cantidad de oraciones para entrenar. Además, solo se trabajo con el grupo de oraciones rotuladas para entrenamiento en la división original de TIMIT, dado que esta presenta mayor cantidad de oraciones por hablante.

Por otro lado, se usaron registros de hablantes femeninos de una misma región, lo cual en nuestro caso incrementa el grado de dificultad del reconocimiento.

2.3 Procesamiento de señales

El objetivo del procesamiento de señales es el de extraer la información relevante de una señal por medio de algún tipo de transformación. Un análisis comúnmente empleado en reconocimiento del hablante es el Mel Cepstrum.

De acuerdo a los modelos usuales, la voz está compuesta de una secuencia de excitación convolucionada con la respuesta impulsó del modelo del sistema vocal **[6]**. Nosotros solamente tenemos acceso a la salida, pero es frecuente encontrar que es deseable eliminar una de las componentes, de tal forma de poder examinar la otra.

La eliminación de una de las dos señales, es en general, un problema muy dificultoso. Sin embargo, hay métodos para resolver este tipo de problemas cuando las señales están combinadas linealmente.

El Cepstrum representan una transformación sobre la señal de voz con dos importantes propiedades:

- Las componentes de la señal son separadas en el Cepstrum.
- Las componentes de la señal son linealmente combinadas en el Cepstrum.

El análisis Cepstra es un caso especial de la clase general de métodos conocidos como procesamiento Homomórficos.

Se han propuesto varios tipos de filtros para las bandas críticas, siendo una de las configuraciones mas usadas el de ventana triangular, en la escala psicoacústica de Mel (*Figura 1*). Un mel es una unidad de medición de la frecuencia percibida de un tono. No se corresponde linealmente con la frecuencia física de un tono, porque el sistema de audición humana aparentemente no percibe el tono en una forma lineal.

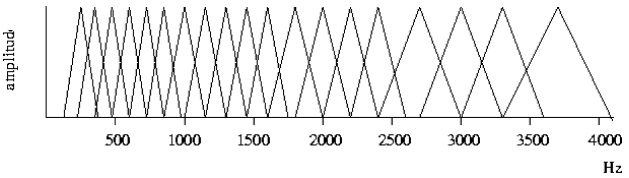

Figura 1: Ventanas triangulares en escala psicoacústica.

Las técnicas anteriores de extracción de características trabajan sobre el espectro de potencia y los coeficientes Cepstrum de la señal. Sin embargo, el espectro de potencia y el Cepstrum no siempre son aconsejables para el reconocimiento de patrones dado que la amplitud y la forma cambian con un simple cambio de micrófono. Una alternativa simple que provee una mayor robustez en los patrones la constituye el Delta Cepstrum. La noción aquí es que la percepción del sonido depende de la diferencia espectral. El Delta Cepstrum calcula la diferencia Cepstral a entre el segmento de voz actual y el anterior para calcular la derivada temporal del Cepstrum.

Algunos sistemas usan solamente el Delta Cepstrum como vector patrón mientras otros usan tanto el Cepstrum y Delta Cepstrum. Este análisis tiene la ventaja de incorporar la información temporal. Por otro lado sufre la desventaja de atenuar información importante en el rango de 1 a 10 Hz.

El procesamiento de señales que se utilizará en el presente artículo lo constituyen los coeficientes Cepstrum y Delta Cepstrum en escala de Mel. Las señales de TIMIT están muestreadas a 16 KHz, luego de varias pruebas los mejores resultados se obtuvieron con una ventana de 512 muestras y un solapamiento de 256, obteniendo luego de la integración por bandas 16 coeficientes de Cepstra adicionándose otros 16 de la derivada temporal. Esto totaliza 32 coeficientes para cada patrón.

2.6 Redes neuronales artificiales

Las RNA intentan simular, al menos parcialmente, la estructura y funciones del cerebro y sistema nervioso de los seres vivos. Una RNA es un sistema de procesamiento de información o señales compuesto por un gran número de elementos simples de procesamiento, llamados *neuronas artificiales* o simplemente *nodos*. Dichos nodos están interconectados por uniones directas llamadas *conexiones* y cooperan para realizar procesamiento en paralelo con el objetivo de resolver una tarea computacional determinada **[7]**.

Si bien los patrones a clasificar son dinámicos, el hecho de utilizar patrones con información de contexto como Delta Cepstrum hace innecesario el uso de *redes neuronales con retardos temporales* (RNRT) **[8,9]**. Por otra parte nuestros experimentos iniciales confirmaron esta hipótesis, por lo que finalmente se utilizó un *Perceptrón multicapa* (PMC). No existe un límite para fijar la cantidad de capas de un PMC, pero se ha demostrado que un PMC con una capa oculta y con un número suficiente de nodos es capaz de solucionar casi cualquier problema. Si se agrega una capa oculta más (PMC-2O), un PMC soluciona cualquier tipo de problemas y en forma más eficiente que con una sola capa oculta.

Para entrenar el PMC se utilizó el algoritmo de retropropagación **[7]**. Las entradas se normalizaron para cada dimensión del patrón en forma independiente. El entrenamiento se detuvo en el pico de generalización medido con respecto al archivo de prueba.

3. Resultados

Los resultados del presente artículo se obtuvieron en base a los siguientes experimentos:

1. Distintas configuraciones de una red única para dos hablantes.
2. Distintas configuraciones de una red única para cinco hablantes.
3. Distintas configuraciones de una red para cada hablante, en un conjunto de cinco hablantes.

En todas las tablas se presentan resultados promedios del porcentaje de frames y oraciones clasificados correctamente para el archivo de entrenamiento (TRN) y prueba (TST).

En la Tabla 1, se presentan los resultados para distintos tipos de redes, correspondientes al experimento 1. Estos se consiguieron variando el número de neuronas en la capa oculta (NNCO), y solo se muestran los que corresponden al mejor desempeño.

La Tabla 2 corresponde a los resultados del experimento 2, los cuales se obtuvieron de forma similar al experimento 1, aunque con cinco hablantes y más tipos de redes ensayadas.

Los resultados de estos experimentos con una sola red y varios hablantes (*Figura 2*) mostraron que el desempeño del clasificador disminuía marcadamene al aumentar el número de clases. Por lo anterior se propuso como alternativa el clasficador de la *Figura 3*, el cual consiste en una red por hablante. Este podría considerarse también como una red más grande, pero parcialmente conectada. En este caso, los archivos de entrenamiento y prueba consisten en oraciones del hablante a identificar, y una selección de oraciones del resto de los hablantes del conjunto.

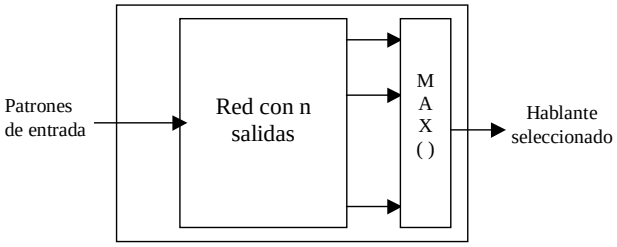

Figura 2: Estructura del clasificador tradicional.

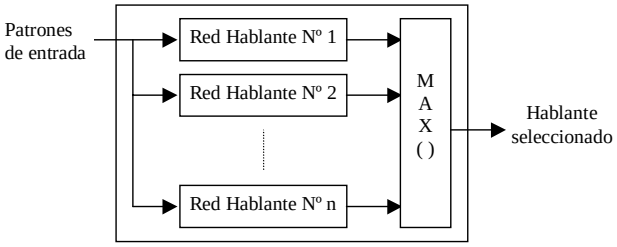

Figura 3: Estructura de clasificador propuesto.

Los resultados del experimento 3 (con el clasificador de la *Figura 3* con *n = 5)*, se presenta en la Tabla 3. La red utilizada fue un PMC-2O, la cual dió los mejores resultados en los experimentos 1 y 2. En la Tabla 4 se muestran los resultados de utilizar las redes ya entrenadas para identificar solo impostores, un grupo de los cuales se encontraban originalmente representados en los archivos de entrenamiento (ICC), y otro grupo que nunca había sido visto por el clasificador.

TABLA 1: RESULTADOS PARA EL EXPERIMENTO Nº 1.

TIPO DE RED	NNCO	% DE FRAMES BIEN CLASIFICADOS		% DE ORACIONES BIEN CLASIFICADAS	
		TRN	TST	TRN	TST
PMC	175	92.34	80.85	100.00	100.00
PMC-2O	125	89.34	85.55	100.00	100.00

TABLA 2: RESULTADOS PARA EL EXPERIMENTO Nº 2.

TIPO DE RED	NNCO	% DE FRAMES BIEN CLASIFICADOS		% DE ORACIONES BIEN CLASIFICADAS	
		TRN	TST	TRN	TST
PMC	100	77.63	50.74	100.00	50.00
PMC-2O	75	76.40	52.62	100.00	70.00
RNRT	200	82.04	51.68	100.00	50.00
RNRT-RO*	100	76.78	51.41	100.00	40.00

* RNRT con retardo en la capa oculta

TABLA 3: RESULTADOS PARA EL EXPERIMENTO Nº 3.

NNCO	% DE FRAMES BIEN CLASIFICADOS		% DE ORACIONES BIEN CLASIFICADAS	
	TRN	TST	TRN	TST
10	84.64	76.88	96.00	100.00
25	84.53	76.75	99.00	100.00
50	87.65	78.11	99.00	100.00
75	86.71	78.09	97.00	100.00
100	83.57	77.76	97.00	96.67

TABLA 4: RESULTADOS PARA EL EXPERIMENTO Nº 3 (IMPOSTORES).

NNCO	% DE FRAMES BIEN CLASIFICADOS		% DE ORACIONES BIEN CLASIFICADAS	
	ICC	ICA	ICC	ICA
10	87.38	76.92	100.00	95.14
25	87.42	76.97	100.00	94.31
50	89.35	76.59	100.00	94.86
75	89.60	79.91	100.00	96.00
100	89.04	78.77	100.00	94.00

4. Discusión

En este trabajo se presenta una alternativa para la implementación de un sistema de identificación automática del hablante independiente del texto basado en la extracción de características acústicas de la señal de voz y la clasificación de estos patrones mediante redes neuronales. Como se puede apreciar de las tablas 3 y 4 el clasificador propuesto no solo presenta el mejor desempeño sino que también es el más robusto. Esto puede deberse a que, a pesar de poseer en total más parámetros, cada red del clasificador se entrena de forma selectiva. La cantidad de neuronas en la capa oculta óptima para estos experimentos es de 75. Estos resultados hacen promisorias su aplicación en un sistema de identificación automática del hablante. Sin embargo, para obtener una conclusión más definitiva al respecto, se deberían realizar experimentos con mayor número de hablantes.

Entre otras alternativas a explorar se encuentran otros tipos de procesamientos, como ser el análisis mediante onditas que ha demostrado ser bastante robusto en tareas de reconocimiento automático del habla **[10]**.

5. Referencias

[1] D. A. Reynolds. R. C. Rose. “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”.. IEE Trans. on Speech and Audio Processing. Vol. 3 Nro. 1 Enero de 1995.

[2] Chi Wei Che. Qiguang Lin. Dong-Suk Yuk. “An HMM Approach to Text-Prompted Speaker Verification ”.

[3] H. Silverman. D. Morgan. “The aplication of dynamic programin to conected speech recognition”. IEEE Acustic. Speech and Signal Processing Magazine. vol. 7. pp 6-25. Julio de 1990.

[4] Chi Wei Che. Q. Lin. D.S. Yuk. “An HMM Approach to Text-Prompted Speaker Verificación”. CAIP Center. Rutgers Uni.. Piscataway. NJ. USA.

[5] Garofolo. Lamel. Fisher. Fiscus. Pallett. Dahlgren. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation*. National Institute of Standards and Technology. February 1993.

[6] J. Deller. J. Proakis. J. Hansen: “Discrete-Time Processing of Speech Signals”. Prentice-Hall. 1987.

[7] Mohamad H. Hassoun. *Fundamentals of Artificial Neural Networks*. The MIT Press. 1995.

[8] J.L. Elman. “Finding structure in time”. *Cognitive Science* 14 (1990) 179-211.

[9] A. Waibel. T. Hanazawa. G. Hinton. K. Shikano. K. Lang; “Phoneme Recognition Using Time-Delay Neural Networks”. *IEEE Trans. ASSP* Vol. 37. No 3 (1989).

[10] H. L. Rufiner, H. M. Torres, “Clasificación de Fonemas Mediante Paquetes de Onditas Orientadas Perceptualmente”, Anales del “1º Congreso Latinoamericano de Ingeniería Biomédia”, Mazatlán, Sinaloa, México, 1998.

Dirección para Correspondencia: Laboratorio de Cibernética – Facultad de Ingeniería (UNER). Ruta 11 Km.10 – Oro Verde (Paraná). Entre Ríos.

Correo electrónico:

Hugo L. Rufiner: lrufiner@arcride.edu.ar

Humberto M. Torres: ipse@santafe.com.ar

sinc(*i*) Laboratory for Signals and Computational Intelligence (http://fich.unl.edu.ar/sinc)
H. M. Torres & H. L. Rufiner; "Identificación Automática del Hablante mediante Redes con Retardos"
Anales del XII Congreso Argentino de Bioingeniería. Vol. 1, pp. 319--322, Jun, 1999.