Mejoras en un Sistema de Reconociendo de Locutor Basado en RNA Mediante Entrenamiento con Normalización de Canal

Carlos E. Vivaracho Pascual Valentín Cardeñoso Payo César González Ferreras

Dpto. de Informática. Universidad de Valladolid
{cevp,valen,cesargf}@infor.uva.es

Resumen. En el presente trabajo vamos a mostrar las mejoras que en el rendimiento de un sistema de reconocimiento de locutor basado en Redes Neuronales Artificiales produce el uso de lo que denominamos *Entrenamiento con Normalización de Canal* (ENC). La mejora en el rendimiento, medido mediante la Tasa de Equierror, veremos que es del 25% con respecto al mejor resultado obtenido si no se usa ENC. Mostraremos, también, el sistema, basado en la técnica anterior, con el que se ha participado en la evaluación estándar propuesta en el International Symposium on Chinese Spoken Language Processing (ISCSLP) 06, ocupando el tercer puesto de los siete sistemas presentados a concurso, muy cerca del segundo. En este sistema se incluye un reconocedor de tipo de canal telefónico usado, que también será descrito.

1. Introducción

Los sistemas de reconocimiento de locutor basados en Redes Neuronales Artificiales (RNA) fueron muy utilizados a principio de los 90 con buenos resultados [1,5]. Sin embargo con el paso del tiempo el uso de este tipo de clasificador ha ido descendiendo. La situación actual muestra un uso muy escaso de este tipo de sistemas. Dentro de éstos podemos destacar los siguientes trabajos recientes [2,3,10].

Nuestro trabajo durante los últimos años se ha centrado en el uso del Perceptrón Multicapa (MLP) como clasificador discriminante. En uno de nuestros primeros estudios comparativos mostramos como el rendimiento de este tipo de sistemas puede ser similar, bajo determinadas condiciones, al de sistemas más populares como los basados en Mezcla de Gaussianas (GMMs) [8]. Desde este trabajo diversas mejoras han sido introducidas [6,7]. En el presente artículo presentamos la última, basada en lo que denominamos *Entrenamiento con Normalización de Canal* (ENC).

Para entrenar un clasificador discriminante, independientemente de su tipo, se necesitan muestras de la clase a reconocer (Clase Objetivo, CO) y representantes del "resto del mundo", es decir, de la Clase No Objetivo (CON). En nuestro caso la CO es el hablante a verificar (hablante cliente) y la CNO es el resto de la población (impostores). El ENC se basa en utilizar como ejemplos de la clase impostores, muestras de voz obtenidas usando el mismo tipo de canal que la muestra del cliente usada para entrenar la red. Veremos como esta forma de entrenar la RNA mejora el rendimiento del sistema de reconocimiento del locutor.

Ahora bien, aparece un nuevo problema: reconocer el tipo de canal usado para obtener la muestra de voz. Para resolver este problema se ha usado también un sistema basado en un MLP discriminante.

El sistema final se ha probado en la evaluación propuesta en el International Symposium on Chinese Spoken Language Processing (ISCSLP) 06. Se ha participado en la tarea de Verificación Automática del Locutor (VAL) independiente de texto "cross channel". Las muestras de voz usadas en la evaluación fueron obtenidas vía teléfono, usando para ello dos tipos de líneas (canales) diferentes: línea terrestre o telefonía fija y telefonía móvil (no se aportó más información acerca del tipo de canal telefónico usado). Dentro de las distintas posibilidades en verificación del locutor se ha participado en la denominada "cross channel", es decir, en la que las muestras de prueba pueden proceder de un canal diferente al usado en la muestra de entrenamiento. El lenguaje de la base de datos usada en la evaluación, como es fácil de deducir por el título del congreso, es el chino. Aunque no se pueden mostrar los resultados obtenidos por los otros participantes, si comentar que nuestro sistema ocupó el tercer lugar, lejos de los resultados del primero, pero muy cerca de los obtenidos por el segundo.

El resto del artículo se organiza como sigue. Empezaremos (apartado 2) con una descripción general del sistema basado en RNA. A continuación (aptdo. 3) mostraremos los experimentos realizados con el conjunto de desarrollo proporcionado por los organizadores de la evaluación, donde se mostrará las ventajas de usar el ENC. Antes de pasar a describir el sistema final presentado a la evaluación y los resultados obtenidos (aptdo. 5), mostraremos el sistema utilizado para reconocimiento del canal (aptdo. 4). Acabaremos con las conclusiones (aptdo. 6).

2. Descripción General del Sistema de VAL Basado en RNA

Se entrena un MLP por cada hablante cliente, usando el algoritmo de retropropagación del error. La arquitectura del MLP es de 3 capas, con 32 neuronas en la capa oculta y 1 en la de salida, siendo todas estas de tipo sigmoideo. La salida deseada para muestras de cliente e impostor se ha fijado en 1.0 y 0.0 respectivamente. Una descripción más detallada se puede encontrar en [6,7].

Aunque las condiciones para aproximar la salida del MLP a una probabilidad a posteriori no se cumplen [4], dado un vector de entrada x_i, perteneciente a una muestra de prueba (muestra de voz) $X=\{x_1, x_2, \ldots, x_M\}$, la salida del λ_C MLP, entrenado para la verificación del hablante C, puede ser vista como la estimación del grado de pertenencia del vector x_i a la clase λ_C. Este valor es representado por $\Gamma(x_i/\lambda_C)$. Siguiendo esta interpretación de la salida del MLP discriminante, aunque no sea una probabilidad real, en vista de su significado y sus valores puede ser tratada como tal. Bajo este enfoque, la pseudo-probabilidad de que una muestra de prueba X pertenezca al hablante cliente C será:

$$\Gamma(\lambda_C / X) = \prod_{i=1}^{M} \Gamma(\lambda_C / x_i) \qquad (1)$$

Como este valor puede ser muy pequeño, para evitar la pérdida de precisión que esto supondría al operar en el ordenador, es aconsejable usar el logaritmo:

$$\log(\Gamma(\lambda_C / X)) = \frac{1}{M}\sum_{i=1}^{M}\log(\Gamma(\lambda_C / x_i)) \qquad (2)$$

El resultado de la ecuación 2 es altamente dependiente de M. Para evitar esto se usa la media, lo que permite además un valor final acotado. El resultado final, *S(X)*, para la muestra de prueba *X* será:

$$S(X) = \frac{1}{M}\sum_{i=1}^{M}\log(\Gamma(\lambda_C / x_i)) \qquad (3)$$

Para mejorar el rendimiento una sencilla pero efectiva regla que hemos denominado R262 [7] es incluida en el cálculo del resultado final por muestra, quedando éste finalmente de la siguiente manera:

$$S_R(X) = \frac{1}{N_R}\sum_{i=1}^{N_R}\log\Gamma(\lambda_C / x_i) \quad \forall x_i / \Gamma(\lambda_C / x_i) \notin (0.2,0.8) \qquad (4)$$

Donde N_R es el número de vectores x_i que verifican la regla en (4).

Dado el tamaño de las bases de datos actuales surge el problema de la selección de las muestras de impostores que serán usadas para entrenar la red. El problema concreto es la gran desproporción que se tiene entre la cantidad de muestras de la clase cliente y de la clase impostor disponibles para entrenar la red, ya que como representantes de la clase impostor podemos usar, en principio, cualquier muestra de voz no proveniente del cliente. Si un MLP se entrena con mucha desproporción entre clases, esta demostrado que tiende a aprender la más numerosa, por lo que su rendimiento es malo.

Tenemos, entonces, que seleccionar de todas las muestras disponibles de la clase impostor, un subconjunto que sea lo más representativo posible. En esta tarea hemos demostrado la eficacia de la técnica que hemos denominado NTIL (Non-Target Incremental Learning), no sólo con RNA [6], si no también con SVM [9]. Dada su eficacia NTIL ha sido usada en el sistema.

Para fijar terminología en lo sucesivo, vamos a denominar al subconjunto de muestras de la clase impostor usadas para entrenar el MLP discriminante Subconjunto Impostor de Entrenamiento (SIE), y al conjunto de muestras de la clase impostor del cual se extrae el subconjunto anterior Conjunto de Muestras de Impostores (CMI).

Por último, para compensar el efecto de que diferentes clasificadores proporcionan diferentes distribuciones de resultados finales, se ha utilizado ZNorm para normalizar esas distribuciones.

3. Pruebas con el Conjunto de Desarrollo. Descripción del ENC

Algunas semanas antes de recibir los datos para la evaluación se recibió un conjunto de desarrollo, con características similares a las del conjunto de evaluación. Este conjunto está compuesto por muestras de 300 hablantes, de cada uno de los cuales se tiene, en promedio, unos 7 ficheros distintos. Para nuestras pruebas este conjunto se dividió de manera aleatoria en los siguientes 3 subconjuntos:

- Subconjunto 1. Compuesto por 100 hablantes que serán usados como clientes. De cada uno de ellos se escogió al azar un fichero que se usó para entrenar su MLP y el resto para prueba. De los otros 99 hablantes del subconjunto se seleccionaron, también al azar, 50 ficheros de 50 hablantes distintos, que fueron usados en prueba como muestras de impostores.
- Subconjunto 2. Consta de otros 100 hablantes, de cada unos de los cuales se seleccionó un fichero al azar, ficheros que componen el CMI. De estas 100 muestras, aproximadamente la mitad provenían de línea terrestre y la otra mitad de móvil, es decir, el tipo de canal estaba equilibrado.
- Subconjunto 3. Compuesto por los otros 100 hablantes. De cada uno de estos se seleccionó una muestra al azar, que fueron usados para aplicar ZNorm.

Para que los resultados fueran más concluyentes, se realizaron pruebas con distintas selecciones de estos 3 subconjuntos. El objetivo principal de estas pruebas fue comprobar el rendimiento del sistema con respecto a:

- La técnica de compensación del canal utilizada en la extracción de características. Se probaron dos diferentes:
  - **Parametrización 1.** De la señal de voz se extraen 19 MFCC (Mel-frecuency Central Coefficients), más sus correspondientes deribadas (19 ΔMFCC). Estos son extraídos de marcos de 32 ms, con solapamiento de 16 ms. Se usó ventana de Hamming. La señal de voz es preenfatizada usando un coeficiente de 0.97. Para compensar el efecto del canal se utiliza la substracción de la media (Cepstral Mean Subtraction, CMN).
  - **Parametrización 2.** A la anterior se le añade la aplicación de la técnica RASTA para potenciar la compensación del efecto canal.
- Usar o no usar ENC:
  - **Entrenamiento con Normalización de Canal.** Consiste en escoger como representantes de la clase impostor para entrenar la red (SIE) sólo muestras del CMI cuyo canal (línea telefónica terrestre o móvil) sea del mismo tipo que el de la muestra del cliente usada para entrenamiento
  - **Entrenamiento sin Normalización de Canal (EsNC).** El tipo de canal de la muestra de entrenamiento del hablante cliente no se tiene en cuenta a la hora de escoger el SIE. Esta es la técnica usada hasta ahora en nuestros sistemas.

En la figura 1 se puede observar el rendimiento del sistema, medido mediante curvas DET, con respecto a las distintas alternativas propuestas anteriomente. En la tabla 1 resumimos este rendimiento utilizando dos puntos significativos de las curvas DET anteriores: el valor óptimo de la Función de Coste de Detección (Detection Cost Function, DCF) y la Tasa de Equierror.(Equal Error Rate, EER). La DCF ha sido calculada como se propone en las normas de la evaluación:

$$C_{DET} = C_{Miss} \cdot P_{Miss} \cdot P_{Target} + C_{FalseAlarm} \cdot P_{FalseAlarm} \cdot (1 - P_{Target}) \quad \textbf{(5)}$$

Donde, $C_{Miss}=10$, $C_{FalseRejection}=1$ y $P_{Target}=0.05$.

Los resultados muestras como con independencia de la parametrización el uso del ENC mejora el rendimiento del sistema, siendo esta mejora superior al usar la *parametrización 1*.

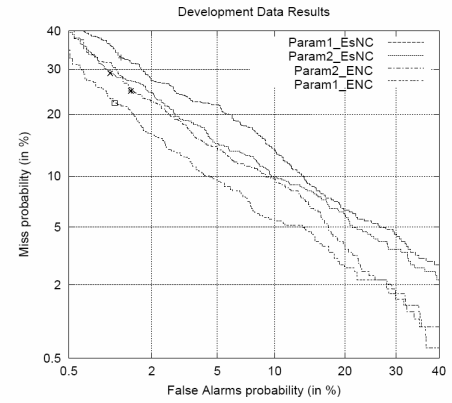

Fig. 1 Resultados obtenidos con el conjunto de desarrollo.

Tabla 1 Rendimiento, medido mediante DCF y EER, obtenido con los datos de desarrollo.

	Par. 1 y ENC	**Par. 1 y EsNC**	**Par. 2 y ENC**	**Par. 2 y EsNC**
DCF	9.1%	14.8%	11.4%	12.0%
EER	7.4%	11.7%	9.7%	10.0%

4. Clasificador de Tipo de Canal

Vista la ventaja de usar el ENC, surge el problema de reconocer el tipo de canal, ya que este dato se proporciona en el conjunto de desarrollo pero no en el de evaluación.

El problema planteado vuelve a ser, al igual que el de verificación del locutor, un problema de clasificación de dos clases, en este caso: línea telefónica terrestre (land) o línea telefónica móvil (cell). La solución planteada fue, por lo tanto, similar a la empleada en es sistema de VAL, es decir, usar un MLP como clasificador discriminante. La arquitectura empleada es la misma que la indicada en el apartado 2.

La extracción de características es también la misma que la mostrada en la sección anterior pero sin utilizar ninguna técnica para compensar el efecto del canal.

Para entrenar el clasificador los 300 hablantes del conjunto de desarrollo se dividieron ahora en dos subconjuntos:

- Subconjunto 1. Compuesto por 250 hablantes. De cada uno de estos se extrae al azar una muestra (fichero) con tipo de canal *land* y otra con tipo de canal *cell*. Estas muestras son usadas para entrenar el clasificador.

- Subconjunto 2. Consta de otros 50 hablantes, cuyas muestras de voz son usadas para probar el sistema.

De los resultados mostrados en la sección anterior está clara la ventaja de usar un ENC, pero, ¿qué pasa si el clasificador de canal falla y entonces para entrenar la red el SIE proviene de una canal distinto al de la muestra de entrenamiento del cliente? En la tabla 2 se pueden ver los resultados. Se ve claramente que el rendimiento del sistema se deteriora, por lo que el coste del error en el sistema de reconocimiento de canal es alto, en términos del rendimiento final del sistema de VAL.

Tabla 2 Rendimiento del sistema con los datos de desarrollo cuando el entrenamiento del MLP se realiza con nomalizacion del canal (columna *ENC*) y cuando el canal de la muestra de entrenamiento del cliente y del SIE son distintos (columna *Distinto Canal Cli/SIE)*.

	ENC	**Distinto Canal Cli/SIE**
DCF	9.1%	19.8%
EER	7.4%	18.1%

Para decrementar la probabilidad de error en la decisión final del sistema de reconocimiento de canal se realizó lo siguiente:

1. Unos 40 clasificadores fueron entrenados con distintas selecciones de los dos subconjuntos anteriores.
2. De esos se escogieron los 4 con mejor rendimiento. La EER obtenida estaba entre el 2% y el 3%.
3. La decisión final sobre el tipo de canal se basó en la decisión individual de cada uno de esos 4 clasificadores: solamente si 3 o los 4 clasificadores clasificaban una muestra como *land* o *cell*, esta muestra era finalmente clasificada como *land* o *cell*, respectivamente; en caso contrario la decisión final era *desconocido*. Como se operó en esta situación se explicará en el siguiente apartado. El umbral de decisión para cada clasificador se fijó en el punto de equierror.

5. Evaluación ISCSLP

5.1 Descripción de los datos

Los datos de la evaluación fueron extraídos de la base de datos del CCC (Chinese Corpus Consortium) CCC-VPR2C2005-100. Es una base de datos telefónica en chino con sólo hombres. De los 1000 hablantes de que consta la base de datos se escogieron ficheros de 800 para la evaluación. El número de pruebas totales es 11787, de las cuales 600 son “auténticas” (corresponden al hablante para el que se entrenó el modelo) y 11187 son de “impostores”. La duración media de las muestras de entrenamiento es de unos 30 segundos. La duración de las muestras de prueba tiene más variabilidad, estando entre los 5 s. de las más cortas y los 40 s. de las más largas.

5.2 Sistemas y Resultados

Debido a la incertidumbre sobre el rendimiento del sistema mostrado en el apartado anterior, se decidió participar con dos sistemas distintos de VAL, lo que además nos permitiría medir de alguna manera ese rendimiento del clasificador de canal en condiciones reales:

- **Sistema 1**. Antes de entrenar la red usamos el clasificador de canal para reconocer el tipo del empleado en la muestra de entrenamiento del cliente. Si la decisión final de este clasificador es *cell* o *land*, entonces la extracción de características de la muestra de voz es la mostrada en la *parametrización 1*, y se usa ENC. Sin embargo, si la decisión del clasificador de canal es *desconocido* no usamos ENC y la parametrización es la 2, ya que ha mostrado un mejor rendimiento en ese caso.
- **Sistema 2**. Si el error del clasificador de canal fuera alto, hemos podido comprobar que el rendimiento del sistema sería malo, por lo que para evitar esta eventualidad en este segundo sistema no se usa ENC y, por la misma razón que la mostrada en el punto anterior, la extracción de características es la mostrada en la *parametrización 2*.

Como CMI se usó el mismo que en el último experimento realizado con el conjunto de desarrollo.

En la figura 2 se puede ver el rendimiento de ambos sistemas mediante curvas DET. Los resultados devueltos por los organizadores de la evaluación sólo muestran el valor óptimo de la DCF para cada sistema participante. Para el *sistema 1* el valor de esta medida de rendimiento es del 10.9 % y para el *sistema 2* del 15.3%. Aunque no está permitido mostrar los resultados de otros participantes sí que queremos comentar que el resultado obtenido con el *sistema 1* es el tercer mejor valor de los 7 participantes, y que, aunque alejado del primero, está muy cerca del segundo.

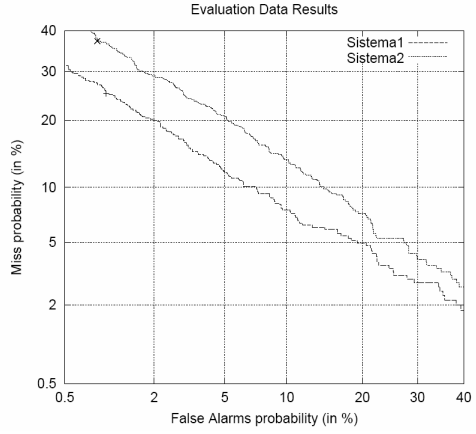

Fig. 2 Resultados obtenidos con los datos de la evaluación

En el fichero de resultados que se enviaba a la organización se podía incluir la decisión sobre cada muestra de prueba. Nosotros incluimos esta información, usando como umbral de decisión el obtenido para el valor óptimo de la DCF en las pruebas

realizadas con el conjunto de desarrollo. La DCF real, obtenida a partir de los errores en la decisión indicada, es de 10.8% para el *sistema 1* y de 15.4% para el *sistema 2*. Como se puede observar estos valores son casi idénticos a los óptimos.

6. Conclusiones

En el presente trabajo además de la tarea de Verificación Automática del Locutor, también se aborda la de reconocimiento del canal de adquisición de la muestra.

Con respecto a esta segunda tarea, podemos observar que la diferencia entre los resultados obtenidos con los datos de desarrollo (donde se conoce el canal) y los obtenidos en la evaluación no es muy alta, lo que nos permite concluir que, aunque el reconocedor de canal se puede mejorar, su rendimiento se puede considerar bueno. Consideramos que la propuesta presentada inicia una vía de trabajo interesante.

Con respecto al sistema de VAL, queremos resaltar el buen rendimiento de la propuesta ENC, su introducción ha supuesto una mejora con respecto a nuestros sistemas anteriores (representados por el *sistema 2* presentado en la evaluación) de un 25%. También queremos resaltar los buenos resultados obtenidos en la evaluación ISCSLP06 usando un sistema basado en RNA. Si bien no ha sido el mejor, está muy bien posicionado, y más teniendo en cuenta que nos enfrentábamos a un lenguaje totalmente nuevo para nosotros.

Por último, creemos que otro resultado interesante es que el valor óptimo de la DCF y el real son casi idénticos. Esto implica que el comportamiento del sistema propuesto parece bastante predecible, lo que significa una propiedad interesante a la hora de implementar sistemas reales.

7. Referencias

1. Artieres, T., Bennani , Y., Gallinari ,P., y Montacie , C..: Connectionist and conventional models for free text talker identi¯cation. Proc. Neuro-Nimes, France, 1991.
2. Ganchev ,Todor, Tasoulis, Dimitris, Vrahatis, Michael N. y Fakotakis, Nikos: Locally recurrent probabilistic neural network for tex-independent speaker verification. Proc. Eurospeech 03, pp. 1673-1676, 2003.
3. Lapidot, Itshak: SOM as likelihood estimator for speaker clustering. Proc. Eurospeech 03, pp. 3001-3004, 2003.
4. Lawrence, Steve, Burns, Ian, Back, Andrew, Tsoi, Ah Chung y Giles, C. Lee: Neural networks classification and prior class probabilities. Lecture Notes in Computer Science, pp. 299-314, 1998.
5. Oglesby, J y Mason, J. S.: Optimization of neural models for speaker identification. Proc. IEEE ICASSP, Vol. S5-1, pp. 261-264. 1991
6. Vivaracho, Carlos E., Ortega-Garcia, Javier, Alonso, Luis, y Moro, Quiliano I.: Extracting the most discriminant subset from a pool of candidates to optimize discriminant classifier training. Lecture Notes in Computer Science/Lecture Notes in Artifcial Intelligence, Foundations of Intelligent Systems, 14th Iternational Symposium ISMIS, No. 2871, pp. 640-645, 2003.

7. Vivaracho, Carlos E., Ortega-Garcia, Javier, Alonso, Luis, y Moro, Quiliano I.: Improving the competitiveness of discriminant neural networks in speaker verification. Proc. Eurospeech, ISSN 1018-4074, pp. 2637-2640, september 2003.
8. Vivaracho-Pascual, C., Ortega-Garcia, J., Alonso-Romero, L. y Moro-Sancho, Q.: A comparative study of MLP-based artificial neural networks in text-independent speaker verification against GMM-based systems. Proc. Eurospeech01, volume 3, pp. 1753-1756, 3-7 September 2001.
9. Vivaracho, Carlos E.: Improving SVM training by means of NTIL when the data sets are imbalanced. A aparecer en LNCS/LNIA 16th International Symposium on Methodologies for Intelligent Systems (ISMIS), 2006.
10. Yegnanarayana , B. y Kishore, S.P.: AANN: an alternative to GMM for pattern recognition. Neural Networks, Vol. 15, No. 3, pp. 459-469, April 2002.