Genesys: Una Nueva Arquitectura Conexionista para el Reconocimiento de Locutores

Angel Garcia Crespo, Belen Ruiz Mezcua, Roberto Rodriguez Galan y Paloma Domingo Garcia

Universidad Carlos III de Madrid

C/ Butarque ,15. 28911-Leganés. Madrid

email:bruiz@inf.uc3m.es

RESUMEN

En los sistemas de reconocimiento de locutores, la obtención de un esquema discriminativo, capaz de aprender las características del locutor que lo identifiquen de una forma única de forma que el sistema sea capaz de diferenciarlo de los demás, constituye uno de los objetivos fundamentales en el diseño de la arquitectura del sistema. Por ello GeneSys es una herramienta eficaz por su fuerte capacidad de discriminación y su capacidad de aprendizaje que absorbe los cambios en las características de la voz que en los locutores se producen con el paso del tiempo.

Introducción

En el mundo del reconocimiento de locutores en sus vertientes de identificación como verificación, existen varios problemas que dificultan su integración en el mundo real. De ellos los más importantes son los problemas relacionados con el ruido del entorno, la variabilidad temporal del habla y la difícil discriminación entre los locutores que componen la base de reconocimiento. En los sistemas de reconocimiento de locutores, la obtención de un esquema discriminativo, capaz de aprender las características del locutor que lo identifiquen de una forma única de forma que el sistema sea capaz de diferenciarlo de los demás, constituye uno de los objetivos fundamentales en el diseño de la arquitectura del sistema. La primera dificultad se encuentra en el establecimiento de un umbral estable, capaz de decidir si las características de un locutor son lo suficientemente parecidas al usuario deseado y lo suficientemente diferentes de los demás. Por ello GeneSys es una herramienta eficaz por su fuerte capacidad de discriminación. Otra de las características intrínsecas de GeneSys es su capacidad de aprendizaje que absorbe los cambios en las características de la voz que en los locutores se producen con el paso del tiempo.

El siguiente dilema es la elección de las características que vamos a utilizar para alimentar dicha herramienta. En el caso de las aplicaciones de sistemas activados por voz existen dos problemas fundamentales: la voz es un sistema variante con el tiempo, y las características del hablante varían con el estado físico, emocional, el entorno y el paso del tiempo. En los sistemas dependientes de texto el mensaje establecido en la generación de los modelos (entrenamiento del sistema) es el mismo que en el test de los mismos. En esta aplicación se ha seleccionado una aplicación dependiente de texto. De los experimentos realizados se han seleccionado como secuencia de entrada a la red las matrices de autocovarianza obtenida de una secuencia de vectores de características espectrales calculados a partir de la locución presentada al sistema.

Análisis Acústico de la Señal

- Para generar un vector de características que sea adecuado para el procesamiento previo a GeNeSys, es preciso verificar que se cumplen las restricciones impuestas por el clasificador, para que este trabaje adecuadamente:
- Que sea estable en el tiempo
- Que represente las características que queremos obtener como identificativas de cada locutor

- Que lo diferencien de otros locutores; es decir que sean suficientemente discriminativas.

En general se precisa tener la información más completa utilizando el menor número de datos posibles. Como características más relevantes de la voz se han seleccionado las siguientes, que se supone define los segmentos de voz sometidos a análisis, diferenciándolos del resto:

- Espectro de la señal de voz. El espectro se define como la envolvente espectral descrita por los coeficientes cepstrum (definidos en el caso de ruido en la escala Mel como se verá en breve). Se supone que la descripción proporcionada por los 10 primeros coeficientes cepstrum en el desarrollo en serie frecuencial tiene un error aceptable.

- Evolución del espectro. Como se ha indicado la señal de voz es variante con el tiempo, por lo tanto es importante conocer como se describe la evolución del espectro definido por los coeficientes cepstrum, calculando los cepstrum diferenciales en una anchura de ventanas de análisis igual a tres. Se define asimismo la evolución espectral con los diez primeros cepstrum diferenciales.

- Energía y energía incremental de la señal de voz.

Una vez obtenida la secuencia de vectores de características formados para la trama de análisis, según se observa en la figura 1, estos sirven de entrenamiento para la generación de las matrices de autocovarianzas. Se han seleccionado estas matrices por representar de una forma unívoca para una secuencia de entrenamiento a cada uno de los locutores que están representados en la base de datos.

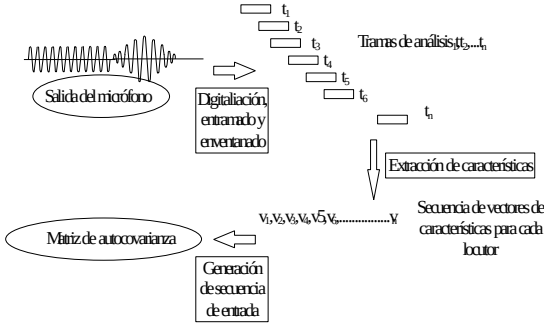

Figura 1: Secuencia de generación de las matrices de autocovarianza.

GeNeSys: Una Arquitectura Conexionista

El reconocimiento de patrones basados en redes neuronales es un procedimiento de clasificación de patrones que intenta realizar el procesamiento de una forma similar al que sigue el cerebro humano. En ese sentido el hombre precisa de una entrada repetitiva de un mismo elemento, que una vez procesado es reconocido como un patrón determinado. En este apartado se ofrece una visión genérica de la arquitectura de la red neuronal denominada GeNeSys.

Consideramos a N como el número de nodos de la Capa de Entrada, I y a M como el número de nodos de la Capa de Salida, O.

1. Inicialización del modelo: La parametrización de las constantes se realiza de acuerdo con las limitaciones siguientes:

   a,b > 0; $0 \le c,d \le 1$; e << 1;
   $K \ge 1$; $0 \le \rho \le 1$; *0 £ θ £ 1;*

2. Los pesos ascendentes y descendentes reciben valores iniciales no nulos:

$$w_{ji}(0) = 1 \qquad w_{ij}(0) < \frac{1}{\sqrt{N}}$$

3. Se inicializa el modelo con el vector nulo, y se da el valor 0 al contador de ciclos.

4. Se incrementa el valor del contador en 1. Se definen seis procesos, p_i, que se aplican al patrón de entrada, P:

   - Proceso 1 (p_1): Adicción del patrón de entrada P y p_4 expandida por la constante *a*.

$$p_{1i} = P_i + ap_{4i}$$

   - Proceso 2 (p_2): Normalización de p_1, ajustada por la constante de normalización *e*.

$$p_{2i} = \frac{p_{1i}}{e + \|p_1\|}$$

   - Proceso 3 (p_3): Adicción funcional de p_2 y p_6 expandida por la constante *b*.

$$p_{3i} = f(p_{2i}) + bf(p_{6i})$$

dónde la forma de la función *f(x)* determina la naturaleza de la mejora de contraste que tiene lugar en I. La elección lógica para esta función podría ser una sigmoide, pero la opción más elemental es la función escalón:

$$f(x) = \begin{cases} 0 & 0 \le x \le q \\ x & x > q \end{cases}$$

en donde *q* es una constante positiva y menor o igual que 1. Otra opción posible para esta función es la sigmoide:

$$f(x) = \begin{cases} \dfrac{2q^2}{(x^2 + q^2)} & 0 \le x \le q \\ x & x > q \end{cases}$$

- Proceso 4 (p_4): Normalización de p_3, ajustada a la constante de normalización *e*.

$$p_{4i} = \frac{p_{3i}}{e + \|p_3\|}$$

- Proceso 5 (p_5): Adicción funcional de p_4 y la expansión funcional del patrón ganador de la capa O.

$$p_{5i} = \begin{cases} p_{4i} + \sum_j g(y_j) w_{Ji} & \textit{si O esta activa} \\ p_{4i} & \textit{si O esta inactiva} \end{cases}$$

- Proceso 6 (p_6): Normalización de p_5, ajustada a la constante de normalización *e*.

5. $$p_{6i} = \frac{p_{5i}}{e + \|p_5\|}$$

6. Se propagan los valores del proceso p_4 hasta la capa O. Se calculan las entradas a O, denominada I_O.

$$I_{Oj} = \sum_{i=1}^{M} p_{4i} w_{ij}$$

Esto permite medir cierta similitud entre el Patrón de Entrada P y las clases de la capa O sin llegar a acceder a dichas clases.

7. Sólo un nodo de O tiene salida no nula, denominado nodo ganador. Este nodo queda definido por la función:

$$g(y_J) = \begin{cases} dI_{OJ} = \max_k \{I_{Ok}\} & \forall k \\ 0 & \textit{en caso contrario} \end{cases}$$

8. Si el valor del contador es 1, se repiten los pasos anteriores con objeto de dar valores a todos los procesos cruzados de la capa I.

9. Calcular la salida de la capa I, denomina O_I, según alguna de las siguientes funciones de semejanza normalizadas. En el caso que nos ocupa se sellecciona la istancia euclidea, aunque si se cambia el patrón de comparación es necesario introducir un cambio en la medida de distancia e incluir el concepto de distancia que mejor se adecue al patrón seleccionado.

Euclidea: $$O_{Ii} = \frac{\sqrt{\sum_i (P_i - w_{Ji})^2}}{N}$$

10. Sin acceder a la clase previamente aprendida, se determina si está se producirá un proceso de aprendizaje. Si no se sobrepasa el *factor de vigilancia*, ρ, entonces se envía una señal de restauración a O y se marcan todos los posibles nodos activos de O como no válidos. Se pone a 1 el contador de ciclos, y se vuelve al paso 2. Si no hay restauración, y el contador de ciclos está a uno, se incrementa el contador de ciclos y se sigue con el paso 11. Si no hay restauración se continua con el siguiente paso.

11. Se modifican los pesos ascendentes de la unidad ganadora de O.

$$w_{iJ} = \frac{p_{4i}}{(1-d)}$$

12. Se modifican los pesos descendentes que provienen de la unidad ganadora de O.

$$w_{Ji} = \begin{cases} P_i(t) & J \textit{ categoria nueva} \\ \dfrac{KP_i(t) + w_{Ji}(t)}{K+1} & J \textit{ categoria existente} \end{cases}$$

13. Se elimina el patrón de entrada, *P*. Se restauran todas las unidas inactivas de O. Se vuelve al paso 2 con un nuevo patrón de entrada *P*.

Arquitectura del sistema

Seleccionado el esquema de clasificación y los parámetros que se utilizaran en la entrada del mismo es

necesario establecer los pasos que se deben seguir para implementar el sistema. Para ello es necesario establecer una plataforma de evaluación en la que se defina el sistema de verificación, tecnología a implementar, la base de trabajo con la que se va a implantar el sistema y el protocolo de experimentación.

Plataforma de evaluación

En ese sentido es necesario acometer dos pasos secuenciales:

En primer lugar se realizará una etapa previa de entrenamiento en la que los modelos asociados a cada locutor serán generados, siendo la salida una base de referencia que contendrá los modelos generados a partir de los datos presentados en la fase de entrenamiento.

Una vez generados los modelos, se procede al reconocimiento propiamente dicho, en el que las características de la voz del usuario que intenta acceder al sistema, son contrastadas con los modelos existentes en la base de referencia o base de datos.

En la figura 2 se resume la secuencia de pasos necesaria para generar los modelos, aplicable para ambas tecnologías, correspondiendo los dos primeros bloques al preproceso de la señal de voz. Este preproceso se lleva a cabo mediante la extracción de características y la obtención de matrices de covarianza. En los modelos basados en GeNesys se hicieron diversas tentativas a la hora de introducir unos pre-modelos a la red. Los que han resultado más eficaces son los basados en las medidas de las matrices de autocovarianza de los vectores de características (formados por elementos de medida espectral y la velocidad de los mismos, así como la medida de la energia y su diferencial) obtenidos de una alocución. Hasta el momento, son estas las características que resultan más interesantes como premodelos para entrenar la red neuronal y categorizar a los patrones obtenidos de cada locutor.

Figura 2: Generación de la base de datos de reconocimiento de locutor

En la fase de reconocimiento tal y como se muestra en la figura 3, los modelos son contrastados con las características de voz del usuario y en función de una medida de semejanza preestablecida, el usuario será aceptado o rechazado en la modalidad de verificación

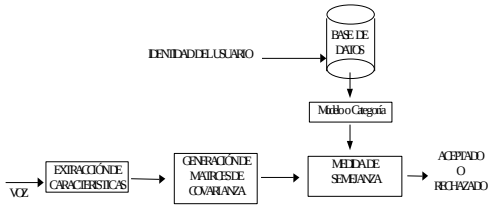

Figura 3: Fase de test en la verificación de locutores

Plataforma de experimentación

Para mejorar las prestaciones del sistema en cuanto a la variabilidad temporal del habla, y permitir a la red que aprenda las variaciones que las características acústicas de los locutores sufren debido al paso del tiempo es preciso dotar al sistema de la información necesaria para que este obtenga sus patrones con este aprendizaje. Para ello se entrena el sistema con la voz procedente de varias sesiones de grabación Para ello es necesario contar con una base de datos multisesión como lo es una base de datos grabada en La universidad Carlos II de Madrid dentro del proyecto ACTS-102 denominado M2VTS (Multimodal Verification Teleservices) que tiene las siguientes características:

- 40 locutores: 20 hombres y 20 mujeres
- Voz microfónica en ambiente de oficina
- 12 sesiones diferentes de grabación separadas entre 5 y 8 días cada una de la anterior. En cada sesión se graban cinco registros o ítems diferentes repartidos como a continuación se indica:
  - Nombre y dirección
  - Fecha y lugar de nacimiento
  - Dígitos del cero al nueve repetidos varias veces
  - Numero de identificación (DNI) pronunciado en dígitos y/o cadenas de ellos
  - Texto libre diferente en cada sesión y para cada usuario

Experimentos realizados

De los experimentos realizados se deduce que la mejor configuración la compone la selección de los dígitos

grabados en cinco sesiones diferentes (un periodo de un mes aproximadamente). Las matrices de autocovarianza obtenidas siguiendo el procedimiento descrito en el apartado de análisis acústico se describe en la siguiente figura:

Figura 4: Muestras de matrices de autocovarianza para un locutor

El entrenamiento consiste en introducir en la red tantas matrices de covarianza de 15 segundos cada una de ellas, como locutores tengamos, es decir se utiliza una única matriz de covarianza para definir a un locutor, esta matriz es producto del análisis de la voz en varias sesiones de grabación.

El entrenamiento se realiza en dos pasos secuenciales:

- Para forzar la creación de una categoría nueva por patrón de entrada, se establece un factor de vigilancia igual al 100%. Así cada patrón se parece a él mismo de forma absoluta y generará una nueva clase categoría. De este modo se consigue generar unas matrices de pesos descendentes que contienen la información de las categorías generadas. Estos pesos son almacenados para su posterior utilización en un fichero. En este paso que se realiza para cada locutor, se obtiene por separado cada uno de los locutores obteniendo un patrón individual para cada uno de ellos. La inclusión de nuevos locutores no afecta a los patrones generados para otros anteriores.
- Cuando todos los pesos son generados y por tanto la información de los modelos contenidos en ellos ha sido obtenida para cada usuario, se introducen en la red cambiando el factor de vigilancia. Los mejores resultados se obtienen cuando este toma un valor igual al 90%. Se actualizan los pesos ascendentes sin modificarse los descendentes. Ambos son almacenados para cada categoría en un fichero y conforman los modelos de cada patrón. En los pesos ascendentes se almacena la discriminación entre todos los locutores y por tanto la inclusión de un nuevo usuario implica realizar esta fase introduciendo los modelos del resto de los usuarios contenidos en la base de datos. También hay que realizar esta operación cuando un usuario es eliminado, para mejorar las prestaciones del sistema.

En la primera etapa de entrenamiento se utilizan las características propias de ART para la creación de categorías y en la segunda la capacidad clasificadora de la red rápida y eficiente.

En el proceso de test se modifica la idea genérica de ART, dado que no se produce aprendizaje adaptativo de patrones. Se utiliza la red pues, tan sólo como herramienta de clasificación y/o rechazo.

Cuando el parecido del patrón de entrada no supera las condiciones del factor de semejanza establecido empíricamente, se clasifica el elemento dentro de una clase basura que contiene todos los errores de clasificación.

En el proceso de test se generan las matrices de autocovarianza de la voz de entrada y se compara con los modelos previamente asignados.

Los experimentos realizados consisten en preparar una serie de patrones de entrenamiento por locutor y entrenar la red de manera independiente para cada uno de ellos, generando una categoría o clase por cada individuo.

Cuando se tienen todas las categorías generadas con la información almacenada en pesos descendentes, se introducen todas las generadas individualmente como patrones de entrada (tantos como locutores haya contenidos en la base de datos). Se hace funcionar la red con un factor de vigilancia del 100%. La finalidad de esta operación reside en obtener los pesos ascendentes que van a identificar a cada una de las categorías, discriminándolas del resto.

Resultados

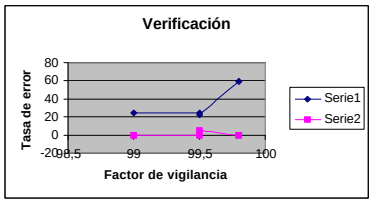

Figura 5: Verificación del locutor en función del factor de vigilancia

La serie 1 describe los resultados de FA y la serie 2 de FR.

En función del tamaño de la base de datos para un factor de vigilancia del 99% se obtiene los resultados descritos en la figura 6 y cuando se tiene 40 locutores de la base de datos los resultados para distinto factores de vigilancia se resumen en la figura 7.

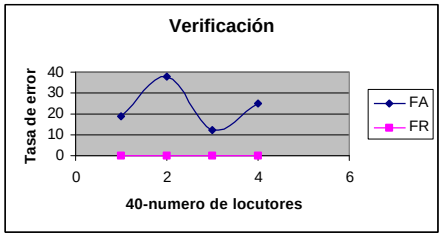

Figura 6: Verificación del locutor en función del número de locutores

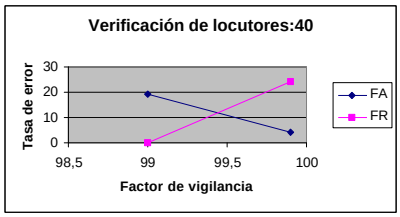

Figura 7: Verificación del locutor en función del factor de vigilancia par 40 locutores

Discusión

Como se deduce de estos experimentos, ART funciona bien cuando se utiliza la modalidad de dependencia de texto y es más robusto frente al falso rechazo que ante la falsa aceptación, es decir es mejor clasificador que discriminados. No obstante estos son experimentos preliminares y con diferentes modificaciones en la estructura y configuración de la red podrían ser mejorados.

Los mejores resultados se obtienen cuando se utilizan 30 patrones de entrada y un factor de vigilancia comprendido entre 99% y 99,5% cuando las sesiones de grabación son diferentes para test y entrenamiento.

Conclusiones

GeNesys se ha mostrado como una nueva técnica de clasificación de patrones, capaz de aprender las variaciones del habla inherentes al paso del tiempo y en el que la inclusión de nuevos locutores permite no disminuir las prestaciones del sistema al ser fácilmente modificables su potente capacidad de discriminación. Se ha mostrado robusto frente a las condiciones del entorno y una herramienta muy potente en la identificación de locutores dado los falsos rechazos obtenidos del 0%. El error obtenido en la falsa aceptación puede reducirse variando el factor de vigilancia en función de los requisitos impuestos por la aplicación.

Este sistema puede implementarse sin requerir grandes prestaciones computacionales. Un prototipo de laboratorio corre en un PC-486, al que hay que añadirle una placa de adquisición de voz que puede ser una Sound-Blaster de 16 bits o compatible.

Referencias

[1] Cáceres-Alonso, P., Rodríguez-Galán, R., and García-Tejedor. A., *Non-Supervised Neural Categorisation of Near-Infrared Spectra. Proceeding of NIR-95*. 7th International Conference on Near-Infrared Spectroscopy. Montreal, Canada. August 6-11, 1995.

[2] Escrihuela, Gerboles y Ruiz (89). *Algoritmica del sistema de reconocimiento de grandes vocabularios*. Documento interno de Alcatel.

[3] Furui & Sondhi *Advances in Speech Signal Processing*. Ed. Marcel Dekker, Inc. 1989.

[4] Lobo, S., García-Tejedor, A.J., Rodríguez-Galán, R., López, L. and García-Crespo, A., *A Simplification of the Theory of Neural Groups Selection for Adaptive Control*. Proceeding de ECAL'95. Granada, Junio 1995.

[5] Rodríguez-Galán, R. and García-Tejedor, A., *ART: an implementation of the new direct access condition*. Proceeding de la International Neural Network Conference. Paris, 1990.

[6] Rodríguez-Galán, R. *Modificaciones al Mecanismo de Aprendizaje de Modelos Neuronales No Supervisados basados en la Teoría de Resonancia Adaptativa. Aplicación al Reconocimiento de Patrones Complejos en Entornos de Producción*. Tesis Doctoral. Dpto. de Matemática Aplicada a la Tecnología de la Información. E.T.S.I.T. Univ. Politécnica de Madrid, 1995.

[7] Ruiz-Mezcua, Gerbolés-Espina, Escrihuela-Langa, Gomez Mena, Veiga. (92) *Reconocimiento de grandes vocabularios independiente del locutor*. URSI92.

[8] Ruiz-Mezcua, Hernadez, Domingo, Rodriguez. *Acceso a servicios multimedia a traves de la voz*. URSI96 Conference.

[9] Ruiz-Mezcua, Lorenzo-Speranzini, García Gómez. *Sistema de verificacion automatica de locutores*. URSI89