TECNIACUSTICA
BARCELONA • 1996

LOQUAX: implementación de un sistema de reconocimiento de locutor en ordenador personal

*Jonathan López, Javier Hernando**

Departament de Teoria de Senyal i Comunicacions
Universitat Politècnica de Catalunya
javier@gps.tsc.upc.es

SUMARIO

Sistematizar el reconocimiento de locutor, es decir, la capacidad de distinguir el propietario o propietaria de un fragmento de voz humana, es un objetivo perseguido desde los inicios del procesado de la señal y enmarcado dentro de la tendencia de conseguir una interacción con las máquinas lo más humana posible. En un principio, plantear un sistema hardware diseñado para el reconocimiento de locutor podía, y puede, resultar muy costoso. Las implementaciones se restringían a simulaciones con computadores potentes o estaciones de trabajo. Hoy en día, gracias a la evolución de la alta tecnología y su implantación en la sociedad, debido principalmente a la popularización del ordenador personal, es posible plantearse un sistema de esta índole que implique una inversión mínima y ofrezca unas prestaciones satisfactorias. Teniendo en cuenta estas consideraciones, este documento presenta una propuesta válida de implementación práctica de un sistema de reconocimiento de locutor: *el proyecto Loquax*.

INTRODUCCION

Reconocer la voz de las personas es una cualidad innata de los seres humanos sin discapacidades auditivas. Es un proceso tan importante como cotidiano. Conversaciones telefónicas, porteros electrónicos,... son ejemplos que forman parte de la vida de mucha gente.

En el campo del procesado digital de voz, se aborda el problema del reconocimiento de locutor desde diferentes estrategias. Las más interesantes son: la *identificación de locutor*, que consiste en pronosticar la identidad de la voz que se presenta y que pertenece necesariamente al grupo de locutores previamente registrados por el sistema reconocedor. La *verificación de locutor*, que consiste en decidir si la voz que se presenta bajo la identidad de un locutor, previamente registrado por el sistema reconocedor, le pertenece realmente. A diferencia de los sistemas de identificación, los sistemas de verificación sólo requieren de la información del locutor que se le presenta. En cambio, es necesario añadirles un dispositivo para poder comunicarles la presunta identidad del locutor. Este dispositivo podría ser un módulo de reconocimiento del habla (sería el caso más cercano al utilizado por las personas, que aprovechamos la voz para revelar nuestra identidad) o bien, en su defecto, teclear su nombre o un número personal [1]. El propósito de esta comunicación es presentar una propuesta de implementación práctica de un sistema de reconocimiento de locutor.

Independientemente de la estrategia escogida, verificación y/o identificación, existen diversos enfoques teóricos de procesado de la voz para implementar un sistema de reconocimiento de locutor. En el apartado siguiente se describen los procedimientos y algoritmos elegidos para el proyecto Loquax y los motivos de dicha elección. En el apartado 3 se detallan las prestaciones principales del sistema implementado y sus diversas fases de elaboración. El siguiente capítulo es un resumen de pruebas y resultados obtenidos y el último apartado resume las principales conclusiones que ha aportado dicho proyecto.

FUNDAMENTOS TEORICOS

El esquema clásico de un sistema de reconocimiento de locutor se muestra en la figura 1. El primer bloque de modelización de la señal de voz tiene la función de extraer a partir de un fragmento de señal de voz

*Este trabajo ha sido financiado por los proyectos TIC95-0884-C04-02 y TIC95-1022-C05-03.

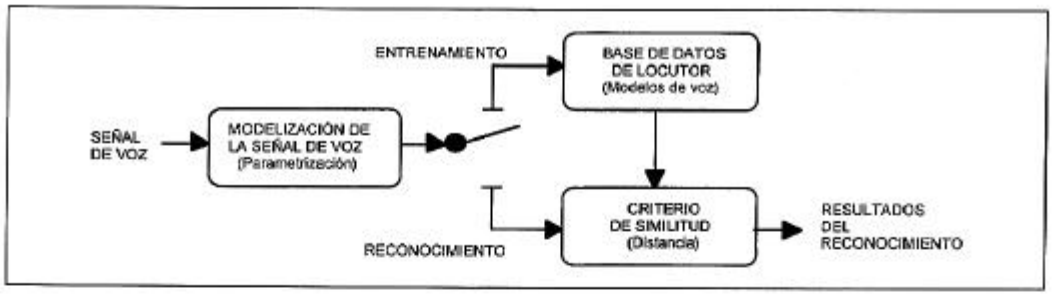

Figura 1: Esquema clásico de un sistema de reconocimiento.

un conjunto de coeficientes que serán la representación del locutor que ha emitido dicho fragmento. Posteriormente a este paso, se distinguen dos procesos básicos: el entrenamiento, es decir, el procedimiento por el cual el sistema "conoce" o "aprende" la voz de uno o varios locutores (la colección de modelos matemáticos que el sistema va "conociendo" se denomina base de datos de locutor), y el propiamente llamado reconocimiento, que produce una serie de resultados comparando los valores recopilados en la base de datos de locutor y los que provienen del locutor cuestionado. Esta comparación se basa en un criterio de similitud, o sea, alguna magnitud o fórmula matemática que dé información sobre el parecido de la voz de dos fragmentos o frases. A este criterio muy frecuentemente se le asocia el concepto de distancia, siendo ésta de valor bajo para voces muy parecidas y de valor alto para voces muy dispares.

Modelización de la señal de voz: matrices de covarianza.

La modelización de la señal de voz que realiza Loquax consiste básicamente en descomponer la señal de voz en tramas, es decir, en pequeños intervalos de 20-30 ms que se solapan parcialmente. Cada una de estas tramas es procesada, calculándose los coeficientes cepstrales correspondientes a un modelo de predicción lineal [2]. Este procesado se complementa con un *filtro de pre-énfasis* para alisar el espectro y un *enventanado* para suavizar el efecto de la descomposición en tramas. La sucesión de vectores cepstrales correspondientes a cada una de las tramas se combinan formando una *matriz* llamada de *covianza* [3], la cual es, finalmente, el modelo de la señal de voz que ha sido procesada. La expresión matemática de esta matriz es la siguiente:

$$M = \sum_{i=1}^{N_T} C_i \cdot C_i^T ,$$

donde N_T es el número de tramas y C_i es el vector i-ésimo de coeficientes cepstrales.

Criterio de similitud: distancia de esfericidad aritmético-armónica

En el caso de modelizar la voz con matrices de covarianza, Bimbot i Mathan [3] recomiendan *la distancia de esfericidad aritmético-armónica*. Dadas dos matrices de covarianza X e Y, la distancia que existe entre ellas se define como:

$$d(X,Y) = \log [A(X,Y) / H(X,Y)],$$

donde A(X,Y) y H(X,Y) son respectivamente la media ritmética y armónica de los valores propios de la matriz XY^{-1}. Esta distancia cumple con los requisitos de toda distancia: es simétrica, no-negativa y además es cero sólo cuando las matrices X e Y son idénticas. Sus propiedades matemáticas ofrecen otras ventajas, como son su gran simplicidad de cálculo y la posibilidad de combinar fácilmente varias matrices de covarianza de un mismo locutor para crear un modelo más fiable.

IMPLEMENTACION DE LOQUAX

Loquax es un proyecto que pretende implementar un sistema completo de reconocimiento de locutor, ofreciendo identificación y verificación, todo ello diseñado para funcionar en entorno Windows, sobre ordenador personal. Se adapta a todo tipo de tarjetas estándar de adquisición de sonido de bajo coste, disponibles en la mayoría de PC´s del mercado actual. Soporta varios formatos de archivos de sonido: el formato WAVE (usado en el entorno Windows) y el formato NIST, usado en colecciones de fragmentos de voz disponibles generalmente en soporte de CD-ROM. El equipo mínimo requerido es un sistema 386 con coprocesador matemático y 2 Mb de memoria RAM, aunque se recomienda un i486 o un Pentium con 8 Mb de RAM. Una unidad lectora de CD-ROM, una tarjeta de sonido estándar de 16 bits y un espacio de disco duro superior a 5 Mb son muy útiles para realizar pruebas sistemáticas de reconocimiento, con bases de datos pregrabadas, con el fin de evaluar el sistema.

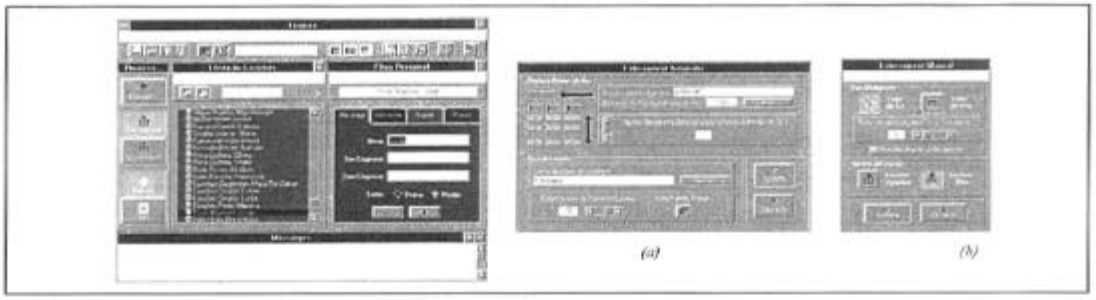

Figura 2: Aspecto de la aplicación Loquax (2ª fase). A la izquieda, visión general. A la derecha, detalles de las cajas de diálogo para construir bases de datos de locutor a partir de colecciones en CD-ROM (a) y sonido directo (b).

Loquax consta de varias fases de elaboración:

- La primera fase consiste en iniciar el proyecto, generando las librerías de procesado y de gestión de datos necesaria para las siguientes fases. El diseño de estos módulos se ha efectuado siguiendo criterios de ingeniería del software, para garantizar su calidad, inteligibilidad y actualizaciones futuras. Los lenguajes de programación han sido: C/C++ para las subrutinas del procesado de la voz, debido a su compilación optimizada, y Visual Basic para la interfície gráfica y la comunicación con las unidades de disco y la memoria, aprovechando su enorme sencillez de programación [4].

- La segunda fase es el desarrollo de una primera aplicación que implementa un sistema de identificación de locutor (ver figura 2). Esta aplicación tiene básicamente dos funcionalidades: por un lado, trabajar con colecciones de fragmentos de voz -bases de datos, típicamente en soporte CD- para poder evaluar la validez y el coste computacional del procesado escogido; por otro lado, poder registrar directamente a los locutores por el propio sistema a través de una tarjeta de sonido, para tener una aproximación real de las prestaciones del sistema.

- la tercera fase consiste en la incorporación de la técnica de verificación de locutor a la aplicación creada en la segunda fase. Ello permitirá tener un sistema completo de reconocimiento de locutor.

- La cuarta y última fase será divulgar el proyecto en universidades, instituciones y congresos especializados, para finalmente llegar a la adaptación de Loquax a cualquier sistema real.

Actualmente se ha finalizado la segunda fase y se trabaja en la tercera y cuarta.

PRUEBAS Y RESULTADOS

Se han realizado estudios estadísticos con 2 bases de datos de locutor compiladas por Texas Instruments y editadas en formato CD-ROM. La primera, llamada TIMIT, consiste en una colección de 630 locutores que registraron 10 frases cada uno de unos 3 segundos de media. El entrenamiento realizado consta de 5 frases por locutor, es decir, el sistema posee información de aproximadamente 15 segundos de voz de cada locutor. Tomando otras 5 frases por cada locutor de un subgrupo de 200 locutores, y realizando una prueba de reconocimiento por cada una de estas frases, se han obtenido tasas de acierto entre el 95.9% (16 coef. LPC y 16 coef. cepstrales) y el 98.6% (30 y 30 respectivamente) [1].

La segunda, llamada TIDIGITS, consiste en una colección de 326 locutores que registraron 77 secuencias de 2 a 11 dígitos, con una media de duración inferior a los 2 segundos. El entrenamiento realizado consta de 4 frases por locutor, es decir, el sistema posee información de unos 7 segundos de voz de cada locutor. Tomando otras 3 frases por cada locutor de un subgrupo de 120 locutores, y repitiendo el proceso anterior, se han obtenido tasas de acierto entre el 89.9% (16 y 16 respectivamente) y el 95.8% (30,30). Si en lugar de efectuar 3 pruebas, se realiza sólo 1 por locutor pero con una frase de unos 5 segundos, los resultados mejoran espectacularmente: entre el 99.1% (16,16) y el 100% (30,30). Estos resultados pueden verse gráficamente en la figura 3. También se han realizado pruebas reales, adquiriendo la voz de los locutores directamente, y se obtienen resultados muy similares.

En cuanto al coste comutacional, el comportamiento del sistema Loquax depende del equipo usado y del número de coeficientes LPC y cepstrales. A nivel de ejemplo, un ordenador basado en un i486-DX2 puede tardar en procesar un fragmento de voz un tiempo comprendido entre 1 y 2,5 veces la duración de dicho fragmento. Un equipo basado en Pentium reduce este cálculo a menos de la mitad. A la velocidad que se mejoran los procesadores hoy en día, se puede afirmar que, sin ninguna modificación, este software funcionará en un futuro muy próximo prácticamente a *tiempo real*.

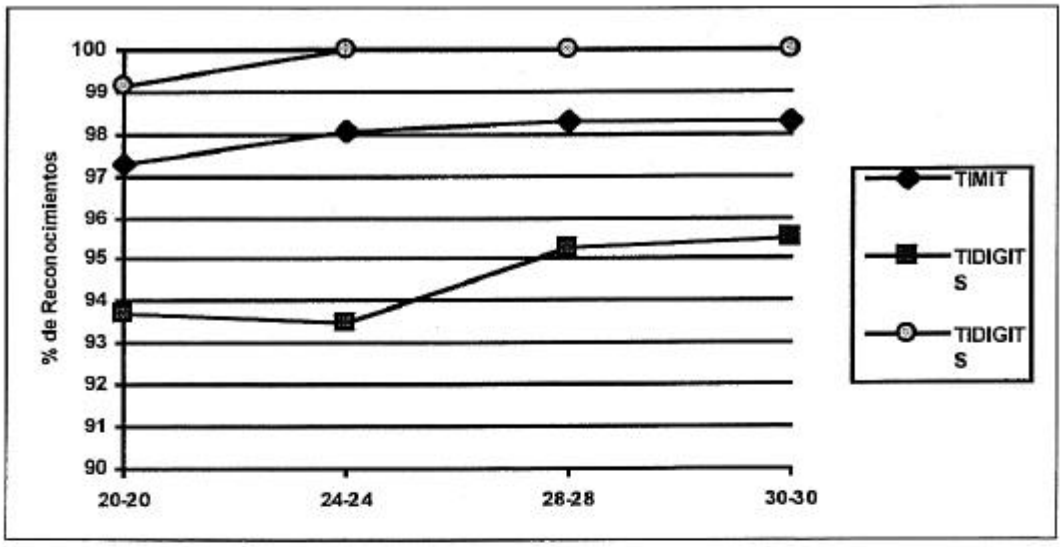

Figura 3: Comparativa de resultados obtenidos con diferentes bases de datos de locutor y parametrizaciones.

El efecto que tiene el número de coeficiente LPC y cepstrales no es tan radical en el tiempo procesado, aunque sí que merece ser tenido en cuenta. El factor más influyente es el número de coeficientes LPC: que este parámetro valga 20 ó 30, significa un incremento del 26% del tiempo de procesado. En cambio, la variación de 20 a 30 del número de coeficientes cepstrales incrementa en menos de un 10% el tiempo de procesado. Por lo tanto, a igualdad de comportamiento, se escogerá una pareja de valores tal que el número de coeficientes LPC sea el menor posible. No olvidemos, sin embargo, que el número de coeficientes cepstrales determina el tamaño de memoria y disco requeridos. Como ejemplo, se necesitan 640Kb de memoria para almacenar 200 matrices de covarianza usando 20 coeficientes cepstrales. Esta cifra se eleva a 1,44 Mb para 30 coeficientes.

CONCLUSIONES

El sistema presentado en este documento ofrece prestaciones de reconocimiento de locutor muy superiores a las que puede ofrecer un ser humano con las mismas condiciones: unos estudios indican que una persona tiene de media un 66% de tasa de reconocimientos para una frase pronunciada por un locutor que pertenece a un grupo de 30 locutores conocidos por el oyente. Este índice está muy por debajo de los indicados en el mencionado apartado de resultados. Así pues, Loquax se presenta como una implementación válida de sistema de reconocimiento de locutor para aplicaciones como control de acceso, servicios telefónicos, etc. Sin embargo, conviene tener una perspectiva realista de este sistema, pues las tasas de acierto pueden bajar debido a condiciones adversas, como pueden ser el ruido ambiental o la falta de cooperación de los locutores. Estudios futuros determinarán su comportamiento bajo dicha condiciones y su posible modificación o adaptación a ellas. A pesar de ello, se puede concluir con un sentimiento optimista en lo que se refiere a la implementación del sistema Loquax de reconocimiento de locutor. La progresión de la última década, tanto a nivel teórico como tecnológico, así nos lo permite. Loquax es una realidad que, aún con su limitaciones, invita a soñar. Miles de aplicaciones le están esperando.

REFERENCIAS

[1] C. Vilagrasa,"*Identificación y Verificación Automática de Locutor*", PFC de ETSETB, 1995

[2] J. W. Picone, "*Signal Modeling Techniques in Spesch Recognition*", Proc. IEEE, Vol. 81, n. 9, pp. 1215-1247, 1993.

[3] F. Bimbot, L. Mathan, "*Text-Free Speaker Recognition Using an Arithmetic-Harmonic Sphericity Measure*", Proc. EUROPEECH'93, Berlin, pp. 169-172, 1993.

[4] R. S. Pressman, "*Ingeniería del Software: un enfoque práctico*", McGraw-Hill, Spain, 1995.