Los bancos de sonidos

RENÉ CARRÉ*

Trad.: M.ª J. Blanco Rodríguez

Aunque actualmente existen numerosas aplicaciones que explotan las técnicas de síntesis o de reconocimiento de la voz, estamos todavía lejos de una verdadera comunicación hombre-máquina que permita dirigir robots por medio de la voz, o bien la consulta, con la ayuda del teléfono, de una base de datos cualquiera.

Con el fin de mejorar la calidad de la voz de síntesis, los esfuerzos deben dirigirse hacia el mismo sintetizador, pero también hacia la orden y la prosodia. En el reconocimiento vocal, la decodificación acústico-fonética representa un escollo que se ha de superar si se quiere ir más allá del reconocimiento de una señal y permitir la comprensión y reproducción de la voz humana.

De este modo, queda claro que el estudio de la lengua es un objetivo prioritario para progresar en este campo. Pero es preciso subrayar, en este sentido, que todos los resultados obtenidos de la investigación sobre una lengua no son utilizables para otra lengua.

Un estudio de estas características es largo y complejo si se tiene la ambición de dominar los aspectos fonéticos, lexicales, sintácticos y semánticos. Nosotros permaneceremos en el nivel de los sonidos, en su descripción acústica y fonética. Los datos correspondientes a esta investigación son numerosos y complejos de interpretar. Por tanto, se hace necesario prever una organización del tipo base de datos para la gestión de los *corpora* y de los

* Director de investigaciones del CNRS. Dirige el Institut de la Comunication Parlée (ICP), que integra ingenieros investigadores del Institut Polytechnique de Grenoble e investigadores fonetistas de la Universidad Sthendhal de Grenoble.

resultados del análisis, banco centralizado para la comunidad científica y con unas características determinadas por consenso. En este sentido, el banco de datos desempeña un papel normativo para los laboratorios fonéticos y permite la ulterior realización de grandes proyectos.

Una encuesta efectuada en Francia en 1982 permitió enumerar veinte *corpora* de sonidos provenientes de siete u ocho laboratorios. Estos *corpora* fueron registrados en diferentes condiciones por lo que se refiere a los locutores, las características de grabación, las técnicas de análisis. Dichos *corpora* no podrían ser utilizados por el resto de la comunidad.

Dos tipos de necesidades

Un banco de datos debe responder principalmente a dos tipos de necesidades:

— Posibilidad de evaluación de los materiales (sistema de reconocimiento, sistema de transmisión, detectores de melodía, etc.) a partir de *corpora* de palabras, de frases adaptadas.

— Posibilidad de realizar investigaciones fundamentales a partir de sonidos en los cuales deben evidenciarse todas las características acústicas: variantes entre interlocutores y en un mismo hablante, variantes regionales, modelización de la producción y percepción de sonidos. Hay que tener en cuenta datos de tipo articulatorio, así como los *corpora* para estudios prosódicos (representación por rayos X, datos electromiográficos, etc.).

Una vez que la comunidad científica haya realizado este primer servicio, se puede pensar en la recuperación por los usuarios de los resultados de análisis obtenidos.

Una primera tarea tiende a etiquetar fonéticamente la señal y a segmentarla, o bien, a especificar los accidentes acústicos característicos. Esta tarea es frecuentemente previa a cualquier otro tratamiento. A continuación, se pueden imaginar numerosos tipos de análisis: detección de índices, detección de formantes, utilización de modelos acústicos, etc.

Una acumulación progresiva de los resultados del análisis a partir del banco de sonidos permite examinar las estadísticas y comenzar, en buenas condiciones, investigaciones sobre el aprendizaje.

Naturalmente, la ampliación del banco de datos acústicos hacia un banco de conocimientos es un objetivo esencial para obtener una mejor respuesta colectiva a las necesidades de síntesis y de reconocimiento.

Los usuarios abastecerán tanto más fácilmente el banco de conocimientos cuanto más se beneficien del conjunto del trabajo de la comunidad. Las posibilidades de utilización de giga-discos (memorias de masa de lectura óptica, análogas a los discos compactos utilizados en música, pero de mayor capacidad: diez elevado a doce bits) permiten considerar la posibilidad de realizar copias completas del banco en cada uno de los laboratorios que los utilicen.

En este contexto, el papel del responsable de un banco de sonidos de estas características consiste en lograr el consenso acerca de la elección de los *corpora*, las normas comunes, las técnicas de análisis normalizadas. Asimismo, es el encargado de organizar la estructura informática de una base de datos relacional, de verificar la calidad de los datos de análisis recuperados, de responder a las necesidades específicas.

Un banco de sonidos de estas características debe poder participar en la siguiente etapa, en el reconocimiento de la voz, o bien relacionarse con la etapa precedente. En síntesis, se trata de la etapa lexical. Se debe, pues, cuidar que las normalizaciones adoptadas en los bancos de sonidos y en los bancos lexicales sean compatibles.

El peligro en la creación de tales bancos de sonidos es olvidar los objetivos de servicio para una comunidad, que consisten, ante todo, en responder de la mejor manera posible a sus necesidades (necesidades de investigación, necesidades industriales). Estos bancos deben, pues, ser manejables y no convertirse en monstruos incontrolables. Es necesario tener en cuenta siempre un compromiso entre agilidad y normalización.

El banco de sonidos del GRECO *Communication parlée*

Se encontrará a continuación el estado del proyecto de investigación *Base de données des sons du français* (BDSON).

La base de datos de sonidos del francés es una de las primeras realizaciones del GRECO (Groupe de Recherches Coordonnées) *Communication parlée*, creado en enero de 1981 por el CNRS.

A pesar de que numerosas dificultades han retrasado demasiado, a nuestro pesar, el inicio de concretización operacional de esta base, es preciso subrayar la importancia del trabajo colectivo realizado, implicando en diversos grados a la mayor parte de los laboratorios franceses.

Un primer grupo de investigadores ha formulado las principales características de la base de datos, la elección de los hablantes, de los primeros *corpora*, la metodología de las grabaciones (Carré y otros). A continuación, un segundo grupo ha estudiado la clasificación fonética y la segmentación (Abry y otros). Actualmente está en desarrollo una reflexión para definir los instrumentos más adecuados para esta clasificación y segmentación. Además, un grupo de trabajo, impulsado por J. Mariani, ha permitido la definición y puesta a punto de programas de análisis, y la definición de unidades informáticas normalizadas.

Actualmente, bajo la responsabilidad de R. Descout, han sido registradas alrededor de treinta y dos horas de habla en el CNET, lo que ha permitido la realización de un estudio de grabación adaptada (Descout y otros, 1986).

Estas grabaciones ahora deberán ser etiquetadas y segmentadas. Además, deberán integrarse en una base de datos relacional que está en proceso de realización en Grenoble.

Elección de hablantes

Son hablantes franceses normales: sobre la base de un conjunto de grabaciones efectuadas en nuestros laboratorios, se ha elegido a los hablantes a partir de los siguientes criterios: nasalidad,

acento, lentitud, claridad, dureza, etc. Un grupo de cinco investigadores ha retenido doce hablantes (seis hombres y seis mujeres) no marcados. Además, se ha elegido a diez hablantes con acento regional, así como diez hablantes considerados como difíciles de reconocer.

Metodología de las grabaciones

Se ha elaborado un gran hardware para permitir al hablante la presentación automática, sobre pantalla, de los elementos del *corpus* y, consecuentemente, poder trabajar sin interrupción en condiciones rigurosamente idénticas. Un «editor de *corpus*» permite, a partir de una simple descripción de la grabación deseada (instrucciones dadas a la pantalla, frases que deben ser pronunciadas para su grabación, períodos de reposo, etc.), realizar con rapidez un programa inmediatamente utilizable para una sesión de grabación. Evidentemente, un hardware de estas características puede ser utilizado para grabaciones de habla en todas las lenguas.

1. Un detector de silencio/habla permite la puesta en marcha automática de la grabación.
2. Los niveles son controlados automáticamente. Una saturación da una señal de error y pide la repetición del elemento presentado.
3. La frecuencia de muestreo, en numeración directa, es de 16 kHz. Además, se efectúa paralelamente una grabación numérica de alta definición en videocasete (Standard Betamax), con un acoplador PCM Sony. Con una frecuencia de muestreo de 44,1 kHz y una conversión en 16 bits, se dispone de esta manera de un «original» de muy alta calidad.
4. Un encabezamiento en cada una de las cintas proporciona informaciones sobre el hablante, sobre el *corpus*, orden de las grabaciones, principio y fin de sonidos o frases, etc.

Un interfaz construido por la Sociedad OROS en Grenoble permite efectuar la conexión entre el conjunto PCM Sony y diversos tipos de ordenadores: PDP11, LSI11, VAX, IBM PC, etc. Este

material permite, además, el paso de la frecuencia de muestreo de 44,1 kHz a un valor cualquiera comprendido entre 5 y 44 kHz. Se puede utilizar una segunda pista para almacenar la información en ASCII.

El GRECO ha equipado doce laboratorios franceses con material de este tipo. Las treinta y dos horas de señales ya grabadas (3,8 gigaoctetos) pueden ser almacenadas en dieciocho casetes (comparables a 320 cintas). Se está evaluando el interés de un material de estas características, que se podría mantener como normalizado.

Los programas perfeccionados en el CNET están actualmente a disposición de los laboratorios franceses para que efectúen sus propias grabaciones. Sus instalaciones podrán obtener el título de GRECO después de inspección. En Francia debían ser instalados tres en 1986.

Primeros *corpora* retenidos

Para la fase preliminar se ha limitado voluntariamente el número de hablantes.

A) *Corpus evaluation*

Hablantes: dieciséis horas y dieciséis minutos:

1. Texto: «La bise et le soleil» (El cierzo y el sol) una vez.
2. CVCV: con vocales /a/, /i/, /u/ y consonantes /p/, /s/, /f/; 54 ítems.
3. Cifras de 0 a 9.100; una lista de aprendizaje más tres listas de reconocimiento; 4 × 100 cifras, una vez; aprendizaje y lista en desorden.
4. Serie de tres cifras; aisladamente, lista de cincuenta series; una lista por aprendizaje (si es necesario) más tres listas de reconocimiento; 4 × 50 series en orden aleatorio.
5. Series de números de dos cifras (encadenados); listas de 50 series; dos veces, orden aleatorio.

6. Serie de números de cinco cifras; listas de 50 series dos veces, orden aleatorio.
7. Lista de cien números entre 0 y 99; dos veces, orden aleatorio.
8. Números de teléfono; 100 ítems.
9. Letras del alfabeto; lista de 27 (con *é* y *e*); aisladamente, en orden; cuatro veces, 4 × 27 = 108.
10. Palabras deletreadas aisladamente; lista de 50 nombres propios; una vez; el nombre aparece en la pantalla más presentación de las letras una después de otra.
11. Palabras deletreadas en cadena; lista de 50 nombres propios; una vez; horizontalmente la palabra encadenada; verticalmente, es decir, deletreándola: *parole*

p
a
r
o
l
e

B) *Corpus acoustique*

Por necesidades de investigación fundamental, los resultados pueden ser utilizados en síntesis, reconocimiento o para estadísticas.

1. *Corpus* CVCV; seis horas, diez veces; con las tres vocales cardinales /a/, /i/, /u/ y las diecisiete consonantes del francés.
2. *Corpus* de grupos consonánticos; dos hombres y dos mujeres, tres veces; con los principales grupos consonánticos del francés.
3. Tests de rimas (palabras aisladas); para las consonantes por pares, por triples; para las vocales, por pares; total 1.330 ítems.
4. Frases fonéticamente equilibradas; 50 frases.
5. Frases para estudio de las nasales; 44 frases.
6. 192 frases con palabras reales formadas con todas las consonantes y todas las vocales del francés.

Las reglas de clasificación y segmentación

Estas reglas han sido objeto de una publicación común (véase Abry y otros, 1985). Una clasificación «amplia» consiste en localizar los centros de las realizaciones fonéticas. Para la clasificación «fin» se han tenido en cuenta dos enfoques: uno se basa en la señal temporal, enteramente manual, y el otro se basa en una representación frecuencial, semiautomática.

En un caso, se retiene alrededor de 100 acontecimientos (principio de sonoridad, fin de la sonoridad, etc.), pudiendo ser interpretados en términos articulatorios. A partir de ellos se deducen segmentos mínimos (microsegmentos) y no mínimos (macrosegmentos). Estas macroclases son del tipo vocal oral, vocal nasal, etc. Una clasificación jerarquizada, que utilice una decena de macroclases, especifica la fase de realización (intensión, tensión y distensión) y los atributos. Estos atributos son del tipo: ruido de explosión, fricción, etc., o bien del nivel superior: vocálico, consonántico, etc.

En el otro caso, un análisis acústico efectúa una segmentación automática a partir de índices espectrales cuyas propiedades son similares a las de los índices utilizados en fonética. Se toman, asimismo, en cuenta los índices contextuales. De esta manera se efectúa un desglose en fonos. A continuación, se explotan los conocimientos fonéticos para efectuar una clasificación.

La puesta en práctica de la clasificación y la segmentación se fundamenta en hardwares experimentados, en particular en Grenoble y Toulouse.

Estructura informática de la base

Se han efectuado varios tests sobre la organización informática de la base de datos. Ha de ser posible efectuar consultas simples, del tipo: «dar los nombres de los hablantes con un acento normalizado». Otras, compuestas, pueden ser del tipo: «dar la lista de realizaciones de un tipo de *corpus* dado, pronunciadas por un hablante dado».

Finalmente, se pueden hacer consultas acerca de ciertos aspec-

tos informáticos de dichas realizaciones: «dar la lista de ítems correspondientes a un *corpus* dado, en todas las realizaciones de todos los hablantes...».

La elección de la máquina y del SGBD no es todavía definitiva. Es cierto que la tarea es inmensa y la cantidad de datos es particularmente imponente: para un primer momento, simplemente para la descripción de los *corpora* ya registrados, es necesaria una memoria de 2 moctetos. Para facilitar el rápido tratamiento de los ítems y responder a la demanda de los laboratorios, parece indispensable, finalmente, un almacenamiento de la señal en gigadiscos.

Normalización de los hardwares y de los materiales

Se han reconocido varios hardwares por el GRECO y han sido objeto de una duplicación comentada. Estos hardwares provenientes de diferentes laboratorios, escritos en FORTRAN, han sido armonizados por una misma persona, bajo la dirección del GRECO. Son los siguientes: FFT, Cepstre, Codificación predictiva, Banco de filtros, Detección de melodía por SIFT, peinado y filtro seguidor.

Una primera reflexión sobre los materiales ha llevado a reconocer los materiales PDP11, LSI11 y SM90. Casi todos los laboratorios disponen actualmente de calculadoras DEC con una norma Qbus y de SM90.

Hoy nos encaminamos hacia una normalización en torno a los PC compatibles IBM. Una placa asociada de tratamiento de señal (fabricada por la firma OROS en Grenoble) permitiría con menor coste el desarrollo de nuestras investigaciones.

Desafíos y perspectivas

El esfuerzo emprendido por el conjunto de la comunidad francesa ha sido muy importante. La coordinación empleada es absolutamente necesaria en este dominio si se quiere disponer de un

conjunto de conocimientos sobre esa lengua y si se quiere participar en las investigaciones internacionales sobre el habla. Pero ese esfuerzo requiere grandes medios para mantener un ritmo de trabajo aceptable y competitivo. Los principios empleados hoy son el resultado de reflexiones de hace entre cinco y diez años. Pensamos que un esfuerzo tal debe ser mantenido en Europa para crear una dinámica nueva y unas subvenciones financieras.

Paralelamente a este proyecto de investigación sobre los sonidos, se está elaborando en el marco del GRECO un banco de datos léxicos para el francés hablado con las variantes fonológicas. Dicho banco está bajo la responsabilidad de G. Perrenou en el CERFIA de Toulouse.

El problema de los bancos de datos no es nuevo. Se plantea en Gran Bretaña, en Suiza, en Italia, etc. En los Estados Unidos se dedican esfuerzos muy importantes al desarrollo de tales bancos. Citamos solamente el MIT, donde con un equipo de una decena de personas y un equipamiento especialmente importante (ocho máquinas LISP y un procesador en matriz), estudia sistemáticamente la lengua inglesa.

Las preocupaciones en Francia son antiguas en este terreno y, por falta de medios, la concreción de nuestro proyecto se retrasa.

Se corre el riesgo de terminar muy pronto en un diálogo hombre-máquina en inglés o japonés, situación que pone en evidencia, más que cualquier otra demostración, la importancia del concepto de industrias de la lengua.

Si los países europeos desean mantener el contacto con los industriales del futuro, deben tomar medidas para hacer su respectiva lengua utilizable y comprensible por la máquina.

Hoy los industriales comienzan a estar sensibilizados con el problema, pero está también en los poderes públicos el facilitar, mediante apoyos financieros consecuentes, los desarrollos necesarios.

Si bien son numerosos los trabajos específicos sobre una lengua concreta, un acuerdo común sobre metodologías entre países europeos, con normalización de algunos materiales para facilitar los intercambios, debería crear una dinámica europea. Por otra parte, el intercambio de resultados sobre nuestras lenguas respectivas puede hacer avanzar la investigación. Por último, los sistemas de diálogo hombre-máquina deberán ser multilingües (en síntesis y en reconocimiento).

Deseo, por mi parte, la puesta en marcha de un proyecto europeo en el dominio de los bancos de sonidos utilizando las estructuras actuales (proyecto ESPRIT o EUREKA, etc.), con una amplia participación de laboratorios de investigación y de industriales.

Bibliografía

Abry, C.; Aytesserre, D.; Barrera, C.; Benoit, C.; Boe, L. J.; Caelen, J.; Caelen Haumont, G.; Rossi, M.; Xock, R., y Vigouroux, N.: *Propositions pour la segmentation et l'étiquetage d'une base des sons du français*, Paris, 14èmes journées d'étude sur la parole, GALF, 1985, págs. 156-162.

Baker, J. M.; Pallett, D. S., y Bridel, J. S.: *Speech recognition performance assessments and available databases*, ICASSP, 1983.

Bommart, T., y Descout, R.: *État d'avancement de la BD-Sons*, Gif-sur-Yvette, Réunion GRECO/industries, 1985.

Carré, R.; Cervantes, O.; Descout, R., y Serignat, J. F.: *Une base de données des sons du français*, Montreal, 11 International Congress on Acoustics, 1986.

Carré, R.; Descout, R.; Eskenazi, M.; Mariani, J., y Rossi, M.: *The French Language Database: Defining, Planning and Recording a Large Database*, San Diego, IEEE ICASSP, 42.11, 1984.

Cervantes, O., y Serignat, J. F.: *Définition et réalisation d'une base de données informatique des sons du français*, Informe interno, 1985.

Doddington, G. R., y Schalk, T. B.: *Speech Recognition: Turning Theory to Practice*, IEEE Spectrum, sept. 1981, págs. 26-32.

Leonard, R. G.: *A Database for Speaker-independant Digit Recognition*, San Diego, IEEE ICASSP, 42.12, 1984.

(*N. B.*) Desde 1984, los encuentros anuales de IEEE dedican una sesión a las bases de datos de sonidos.