Battaner, E., Gil, J., Marrero, V., Llisterri, J., Carbó, C., Machuca, M. J., . . . Ríos, A. (2003). VILE: Estudio acústico de la variación inter e intralocutor en español. In *SEAF 2003. Actas del II Congreso de la Sociedad Española de Acústica Forense*. (pp. 59-70). Barcelona: Sociedad Española de Acústica Forense.

http://liceu.uab.cat/~joaquim/phonetics/VILE/VILE_SEAF03.pdf

# VILE: Estudio acústico de la variación inter e intralocutor en español

*Elena Battaner, Juana Gil, Victoria Marrero*
*Departamento de Lengua Española y Lingüística General, Universidad Nacional de Educación a Distancia*

*Joaquim Llisterri, Carme Carbó, María Jesús Machuca, Carme de la Mota, Antonio Ríos*
*Departamento de Filología Española, Universidad Autónoma de Barcelona*

*vile@liceu.uab.es*

***RESUMEN***

El proyecto VILE tiene como objeto el estudio fonético acústico de la variación inter e intralocutor en español para aplicar sus resultados al reconocimiento automático de hablantes y a la práctica de la fonética forense. En este trabajo se presentan los resultados de una primera etapa, consistente en una revisión de las publicaciones que abordan los parámetros acústicos relevantes para la identificación del hablante y en un estudio de los corpus orales existentes en español a partir de los cuales se puede constituir el conjunto de datos que serán analizados en las siguientes fases del proyecto.

***ABSTRACT***

The aim of the VILE project is the acoustic phonetic analysis of inter and intra-speaker variation in Spanish with particular emphasis in obtaining useful results for automatic speaker identification and for forensic phonetic practices. The first phase of the project – presented in this paper - has consisted in a review of the literature dealing with the acoustic parameters which are relevant for speaker identification, and in an analysis of existing spoken corpora in Spanish from which the data analysed in the project will be taken.

## *1. EL PROYECTO VILE*

VILE (Estudio acústico de la variación inter e intralocutor en español) es un proyecto de investigación[1] cuya finalidad es el estudio acústico de la variación fonética en español, con objeto de aplicar los resultados al reconocimiento automático de locutor y a la práctica de la fonética forense. VILE pretende alcanzar tres objetivos: (i) caracterizar acústicamente los elementos segmentales y suprasegmentales que contribuyen a establecer la individualidad de un hablante frente a aquellos que son comunes a un estilo de habla, una variedad geográfica o social, o una lengua; (ii) obtener el conocimiento fonético necesario para la mejora de los sistemas de reconocimiento, identificación o verificación automáticas del locutor; y (iii) dotar a los especialistas en fonética forense de nuevos datos acústicos que permitan comparar, con un mayor grado de certeza, locutores dubitados e indubitados.

La primera etapa del proyecto se ha centrado en la delimitación de los fenómenos fonéticos objeto de interés, tal como se presenta a continuación en el apartado 2. Para ello, se ha realizado una revisión bibliográfica en tres ámbitos, considerando tanto los aspectos segmentales como suprasegmentales: estudios de fonética acústica del español, estudios sobre

[1] Financiado por el Ministerio de Ciencia y Tecnología (BFF2001-2551, 2001-2004). Puede encontrarse más información sobre el proyecto en `http://liceu.uab.es/~joaquim/VILE.html`

reconocimiento de locutor basado en parámetros fonéticos y estudios de fonética forense. Teniendo en cuenta la necesidad de reutilización de recursos, se ha llevado también a cabo un análisis de los corpus orales existentes en español, valorando su utilidad en relación con los objetivos del proyecto; las conclusiones de esta evaluación se exponen en el apartado 3 del trabajo.

Las próximas fases de VILE se centrarán en el análisis acústico de los fenómenos fonéticos seleccionados en una muestra de materiales extraída de los corpus que finalmente se han considerado más adecuados a los objetivos del proyecto. Los resultados obtenidos se estudiarán desde la perspectiva de la variación, tanto entre hablantes (variación interlocutor) como en un mismo hablante (variación intralocutor), considerando la relación entre ambos tipos de variabilidad y su relevancia para el reconocimiento automático de locutor y para la fonética forense.

## *2. DETERMINACIÓN DE LOS FENÓMENOS FONÉTICOS ANALIZADOS*

### *2.1 Parámetros relevantes para el estudio de la individualidad de la voz*

En el conjunto de la investigación en torno a la identificación del hablante se pueden diferenciar tres tipos de estudios:

(a) los centrados en el reconocimiento visual de los espectrogramas (por ej., Tosi *et alii* (1972)); (b) los que tratan el reconocimiento perceptivo del hablante (por ej., Pollack *et alii* (1954), Compton (1963), Stevens *et alii* (1968), Hollien *et alii* (1982), Kuwabara y Takagi (1991), Kreiman y Papcum (1991), Pisoni (1993)); y (c) los que se ocupan del reconocimiento automático del habla, que son los más frecuentes y constituyen la mayor parte de los consultados para este proyecto (por ej. los estudios de Atal (1972), Wolf (1972), etc.).

Conviene quizá empezar con una reflexión crítica a menudo repetida en la bibliografía (*vid*. por ejemplo Künzel 1995) acerca de los trabajos comprendidos en el grupo a), es decir, aquellos basados en el uso para la comparación de voces de una técnica presuntamente objetiva consistente en la interpretación visual de espectrogramas de banda ancha, técnica que, como precisa Künzel, sigue empleándose en España. Es la llamada técnica de “voiceprint”.

Los rasgos más destacables que se suelen emplear al aplicarla son -aunque ninguno de sus valedores los lista- el ancho de banda de los formantes, sus frecuencias centrales o la composición espectral de las fricativas y de las oclusivas. Se juzga la similitud visual de estas características sobre la asunción de que la diferencia interlocutor es mayor que la intralocutor, lo cual no siempre se constata en los espectrogramas. Se ha demostrado que el margen de error de esta técnica es muy elevado, y que en realidad lo único que hace es desplazar el alto grado de subjetividad que encierran los juicios auditivos o perceptivos al campo visual. Por estas razones no nos hemos detenido en los trabajos del grupo a).

Nos hemos centrado en los trabajos incluidos en los puntos b) y c), es decir, en el análisis de otros procedimientos en los que la influencia del factor humano, aunque complementario, se limite y en los que se consideren parámetros acústicos lo más objetivos y lo más específicos del hablante posible. Doddington (1985) establece una primera división entre los parámetros *de alto nivel de información*, como los referidos al dialecto, estilo, etc., y los de *bajo nivel de información*, como amplitud espectral, frecuencia del tono de voz, frecuencias formánticas, y otros rasgos acústicos. El primer grupo se corresponde con las denominadas *dimensiones socio / psicológicas* (Kuwabara y Sagisaka, 1995), esto es, todos los rasgos dependientes de factores sociales, económicos, geográficos, educativos, psicológicos, físicos transitorios, sexuales o lingüísticos; el segundo grupo, en cambio, se corresponde con las *dimensiones fisiológicas*, que son las que podemos abordar en esta investigación. K.N. Stevens (1971) destacó, entre aquellas

especialmente susceptibles de emplearse en la identificación y discriminación entre hablantes, la frecuencia media del tono ($F_0$) y la *forma de la onda glotal*, que es muy diferente de hablante a hablante como aspectos referidos a la fuente, y una serie de características relacionadas con los resonadores que se enumeran a continuación.

*Frecuencias formánticas*. Si se consideran los valores medios de las frecuencias formánticas en un número de vocales suficientemente amplio, se obtiene un indicio de la longitud media del tracto vocal del hablante. Para este propósito es particularmente útil el valor medio del F3, puesto que este no cambia de modo notable de vocal a vocal y proporciona una indicación más precisa de la longitud del tracto vocal que el F1 y el F2. A medida que aumenta la longitud del tracto vocal, disminuye la frecuencia del formante.

En cuanto a la *anchura de los formantes*, resulta especialmente interesante comparar los de la vocal [i], que parecen diferir poco en el caso de un solo hablante y sin embargo presentan claras diferencias en el caso de varios locutores.

*Sonidos Turbulentos*. En el caso de la [s], se producen algunas diferencias intralocutor, pero claramente menos marcadas que las que se dan entre distintos locutores. Las resonancias de alta frecuencia del tracto vocal que se ven excitadas en la producción de un sonido turbulento como es este dependen de la forma de las cavidades anteriores a la constricción y del modo en el que la lengua y el paladar se disponen en la parte inmediatamente posterior a la constricción.

*Consonantes nasales*. Lo mismo que en el caso anterior, parecen presentar diferencias espectrales más marcadas entre hablantes que intralocutor, si bien el propio Stevens reconoce que las muestras analizadas son muy próximas en el tiempo. Probablemente todas estas características se vieran alteradas con muestras más distanciadas temporalmente.

Hollien (1990) y (1991) mantiene algunos de estos, pero añade otros rasgos, que a su juicio presentan una alta probabilidad de ser decisivos en la discriminación entre hablantes:

*El Espectro a Largo Plazo*. Especialmente útil con datos normalizados obtenidos en laboratorio; muy resistente a los efectos del estrés sobre el habla; en su sistema es el resultado del análisis de 40 parámetros extraídos de la señal.

*El Vector de los formantes vocálicos*. Es un parámetro muy importante para la identificación de los sujetos, porque el tracto vocal individual presenta estabilidad, y porque estos rasgos son muy resistentes a la distorsión y a las interferencias. Todavía resulta interesante el trabajo de Ladefoged y Broadbent (1957) en el que se defiende este parámetro. Hollien, tras revisar la bibliografía sobre la cuestión, elige dos parámetros para configurar su vector: las frecuencias centrales de los *tres primeros formantes* (que parecen ser muy reveladoras, especialmente si se estudian al menos tres vocales, [a, i, u] y la sílaba [na]) y la distancia entre estos tres primeros formantes (F1/F2, F2/F3), que no puede ser alterada a voluntad (*cf.* Tosi *et alii*, 1972).

*El Vector temporal*. Sobre este último se ha trabajado poco (*vid.*, por ejemplo, Johnson *et alii*, 1984), pero en buena lógica pueden ser factores muy importantes para la identificación. Las medidas empleadas para este vector incluyen: (i) el tiempo total de habla, definido como el tiempo en milisegundos que lleva producir una emisión de un conjunto dado de sílabas; (ii) la proporción del tiempo de habla, definido como la medida del tiempo total durante el cual existe energía acústica en una emisión; (iii) la proporción de los intervalos de silencio; (iv) la velocidad de habla, medida de las sílabas completadas durante un periodo de tiempo fijado; y (v), la ratio de la duración consonante / vocal, esto es, la relación entre el tiempo destinado a la producción de la consonante y el destinado a la vocal en una emisión dada de CV.

*El Vector del $F_0$.* En el trabajo de Hollien y colaboradores, este vector implica la medida de 30 parámetros diferentes, por lo que los autores sostienen que los resultados son más fiables que los ofrecidos por estudios anteriores.

Todos estos vectores, reunidos, proporcionan un 'perfil' complejo del hablante, basado en datos naturales, esto es, extraídos de la señal hablada.

En esta misma línea, otros trabajos como Atal (1972), Karlsson (1988), Eskenazi *et alii* (1990), Kuwabara y Takagi (1991) y Kuwabara y Sagisaka (1995) mencionan aproximadamente los mismos parámetros como responsables de la individualidad de la voz. El resumen de todos estos estudios y de los que en ellos se citan podría ser este:

(1) Parámetros referidos a la fuente: *valor medio de la $F_0$, contorno tonal, forma de la onda glotal* y *fluctuación de la $F_0$.*

(2) Parámetros referidos a los resonadores: *frecuencias formánticas; anchura de los formantes; trayectorias de los formantes; distancias y ratios entre formantes; LTAS (Long Term Averaged Spectrum); sonidos turbulentos; consonantes nasales; Efectos coarticulatorios (vocales, nasales y líquidas).*

(3) Variables temporales: *tiempo total de habla; proporción de habla y de silencios; velocidad del habla*

A la hora de seleccionar los parámetros para el estudio, pueden tenerse en cuenta algunas de las reflexiones de Wolf (1972) acerca de cuáles serían criterios de decisión válidos:

(a) Deberían ser parámetros presentes *natural* y frecuentemente en el habla normal. Los efectos coarticulatorios, en cuanto que son, en cierto grado al menos, 'aprendidos', no se tendrían en cuenta (*cf.* para una opinión contraria, Su *et alii* (1974)).

(b) Han de ser fáciles de medir.

(c) Deben tener la mayor variabilidad posible interhablantes y la menor posible intrahablante. Tanto la forma de la onda glotal como los sonidos turbulentos parecen reunir esas condiciones.

(d) No deberían variar mucho con el tiempo ni verse afectados por las condiciones psico- o fisiológicas del hablante. Las nasales quedarían eliminadas de acuerdo con este último criterio (*cf.* sin embargo Wolf (1972) y Glenn y Kleiner (1968)), quienes consideran muy informativo el análisis de las nasales). En el habla espontánea, un rasgo especialmente sensible al estrés experimentado por el hablante es la $F_0$, y uno muy resistente es el LTAS (Pittman 1987, Hollien 1990).

(e) Han de ser resistentes al posible ruido ambiental y no han de verse afectados por las condiciones de la transmisión. Los valores medios de los formantes vocálicos (del F3 –que proporciona indicios sobre la longitud del tracto vocal del hablante- , del F2 y del F1) y su anchura son parámetros muy válidos en este sentido, puesto que son muy resistentes a la distorsión y a las interferencias. Se aconseja estudiar las vocales extremas [a i u]. Por otra parte, se señala que los valores del F1 y F2 de las vocales extremas /i,u,a/ son los más estables y menos sensibles al contexto (Stevens y House 1963).

(f) Finalmente, no pueden ser fácilmente modificables por la mera voluntad del hablante, es decir, deben ser resistentes a los intentos de disimular la voz. La distancia entre los tres primeros formantes no puede ser alterada a voluntad (F1 / F2, F2 / F3).

### *2.2 La importancia relativa de los distintos parámetros*

En este punto es donde más desacuerdo existe, en función de los resultados obtenidos por cada autor en sus experimentos. Las conclusiones sobre cuál es el parámetro prioritario para la individualidad de una voz varían, aunque los diversos autores consultados tratan de asignar un orden jerárquico a los distintos índices. El resultado final sería el siguiente:

(a) Prioridad del $F_0$: Compton 1963), Wolf (1972), Matsumoto *et alii* (1973), Brown (1981), van Dommelen (1987).

(b) Prioridad del LTAS: Bordone-Sacerdote y Sacerdote (1969), Doherty (1976), Hollien y Majewski (1977), Furui (1978 y 1986), Pittam (1987), Gelfer *et alii* (1989).

(c) Prioridad de la estructura espectral, bien sea de las frecuencias formánticas absolutas: Shearme y Colmes (1959), Miller (1964), Itoh y Saito (1982), Carrell (1984) Kuwabara y Ohgushi (1987), Kuwabara y Takagi (1991), o bien sea de las trayectorias formánticas: Ingram *et alii* (1996). Por lo que se refiere a qué formantes o qué distancias entre formantes son las más informativas, tampoco hay coincidencia en las posturas:

(c1) Hollien (1990): F1, F2 y F3, y distancias entre F1-F2 y F2-F3; Kuwabara y Tagaki (1991): F1, F2 y F3; Kreiman y Papcum (1991): F1, F2 y F3 y distancias F2-F1.

(c2) Furui (1986) y Ramón *et alii* (2000): La información más representativa del hablante estará localizada entre los 2.5 KHz. y los 3.5 KHz.

(d) Misma prioridad para la $F_0$ y la estructura formántica: La Riviere (1975).

(e) Variables temporales: Pruzansky (1963), Wolf (1972), Doherty y Hollien (1978), Brown (1981), Johnson *et alii* (1984).

(f) Contorno tonal: Atal (1972), van Dommelen (1997).

(g) No es factible establecer una prioridad, la importancia de cada parámetro puede diferir de hablante a hablante y depende también de la naturaleza de las muestras: Gobl (1989) y Kuwabara y Sagisaka (1995).

### *2.3 Conclusiones*

Tras esta primera revisión de la bibliografía, pueden extraerse las siguientes conclusiones provisionales:

(1) En el material revisado no se han encontrado muchos títulos recientes, pues abundan los publicados en los años setenta y ochenta, que se asumen como punto de partida –sin cuestionarse- en los estudios más actuales.

(2) La falta de resultados coincidentes y concluyentes acerca de cuál sea el parámetro más decisivo para el reconocimiento del hablante puede deberse a los enfoques metodológicos empleados, que son muy diferentes. En cualquier caso, parece difícil establecer una jerarquía absoluta entre los parámetros. Varios autores apuntan la interdependencia entre los índices, cuya prioridad relativa dependería, asimismo, del hablante. Por ejemplo, los oyentes pueden tomar como clave primaria para el reconocimiento de un hablante A un tono bajo, y sin embargo apoyarse en la estructura formántica para el reconocimiento de un hablante B (van Dommelen 1987). Esto es, todos los rasgos mencionados conllevan un cierto grado de información sobre las características del hablante y son potencialmente válidos para la tarea de reconocimiento.

(3) Los estudios más recientes se centran sobre todo en el aspecto metodológico, esto es, en los diferentes sistemas (paramétricos o no paramétricos) de manipulación y control de la cualidad de la voz. Son numerosos los trabajos realizados por especialistas en el campo de la telecomunicación y escasean, en cambio, los estudios de naturaleza puramente fonética,

especialmente los de fonética articulatoria (diferencias entre hablantes en el control y coordinación de las variables articulatorias y sus correlatos acústicos).

(4) Resulta muy evidente que la variabilidad del hablante no se ha investigado en la misma medida que los aspectos invariantes de la producción del habla. Por lo tanto, en esta investigación se deberá trabajar a menudo con bibliografía cuyo objetivo es el contrario al que se ha definido para el proyecto.

(5) En cuanto al número de locutores del que se ha partido para realizar los distintos estudios consultados, es muy variable: desde un mínimo de 8 hasta un máximo de 40 voces distintas.

## *3. SELECCIÓN DEL CORPUS DE ANÁLISIS*

Un requisito esencial para cumplir con los objetivos del proyecto VILE es el acceso a las bases de datos de voz realizadas con anterioridad sobre nuestra lengua. En el momento actual, y dejando al margen los corpus privados, se cuenta con una gran cantidad de recursos orales, ficheros de voz obtenidos en condiciones controladas, tanto desde el punto de vista electrónico, como acústico y fónico. Se trata de un conjunto de datos obtenidos casi siempre al amparo de amplios proyectos de investigación financiados por entidades públicas. Su reutilización parece pues no sólo conveniente, sino también necesaria. Sólo si éstos corpus no llegaran a cubrir aspectos que el estudio bibliográfico (apartado 2) revelara como esenciales para el estudio de la variación inter e intrahablante, el equipo de investigación de este proyecto se plantearía la obtención de una base de datos propia.

### *3.1 Corpus disponibles en español*

Han sido documentadas numerosas bases de datos orales en español, unas generales, otras específicas, algunas fácilmente accesibles, y otras más restringidas. Teniendo en cuenta múltiples factores se han seleccionado, en función de los intereses de VILE, las que se presentan a continuación

ALBAYZÍN es la gran base de datos oral desarrollada en España para reconocimiento y procesado del habla. Se llevó a cabo entre 1992 y 1998, con financiación de la CICYT, por un consorcio que agrupaba, bajo la coordinación de la Universidad Politécnica de Cataluña, a los principales grupos de investigación en tecnología del habla del país.

| | |
|---|---|
| Locutores | 152 hombres, 152 mujeres. |
| Canal | Grabación microfónica en cámara aislada |
| Tareas | Lectura.<br>- Corpus fonético: 700 frases fonéticamente equilibradas, en dos subcorpus, uno de aprendizaje con 4 locutores y otro de prueba con 40 locutores<br>- Corpus de aplicación: 3900 frases sobre datos geográficos<br>- Corpus Lombard: 50 frases de las del corpus fonético leídas a alta intensidad mientras el locutor es sometido a un ruido intenso por los auriculares. |
| Ficheros de voz | 15.600 grabaciones de frases fonéticamente equilibradas |
| Referencias | Moreno *et al.* (1993); Casacuberta *et al*. (1992); Díaz Verdejo *et al.* (1998) |

EUROM1. Se describe habitualmente como la primera base de datos oral europea realmente multilingüe. Gracias al proyecto Esprit SAM-A, el español se incorporó a este gran proyecto de la UE, grabada en las mismas condiciones, con el mismo número de sujetos y un corpus equivalente para once lenguas de nuestro entorno.

| | |
|---|---|
| Locutores | 30 hombres<br>30 mujeres |
| Canal | grabación en cámara anecoica |
| Tareas | Lectura:<br>- dígitos aislados y concatenados<br>- 82 logatomos: C*i/a/u*C*a*, CC*ala*, aisladas y en contexto<br>- 10 palabras aisladas (parte de la frase portadora de los logatomos):<br>*pon, siempre, lejos, pones, aquel, quieto, di, igual, orando, dijo*.<br>- 40 párrafos con cinco oraciones cada uno (comunes), con bastante variación suprasegmental (interrogativas, exclamativas, enumeraciones...)<br>- 50 frases específicas para cada lengua; las del español prestan especial atención a los fonemas palatales, la vibrante múltiple, las fricativas anteriores, secuencias vocálicas, grupos consonánticos, etc. |
| Ficheros de voz | Sin considerar las tareas de dígitos, 770 grabaciones, entre pseudo-palabras, frases y párrafos. |
| Referencias | Llisterri *et al.* (1993) |

MULTEXT. Se trata de un subconjunto de EUROM al que se ha prestado una especial atención en el nivel prosódico. Contiene el $F_0$ original y estilizado, con transcripción y codificación sobre 15 párrafos seleccionados de los 40 que ofrece EUROM

| | |
|---|---|
| Locutores | 5 hombres, 5 mujeres |
| Canal | grabación en cámara anecoica |
| Tareas | Lectura:<br>- quince párrafos de EUROM1 por hablante; duración total: 52:21'.<br>Extraído el F0, estilizado y resintetizado. Ofrece transcripción ortográfica, forma de onda, codificación simbólica del F0 (7 categorías), F0 original y curva estilizada superpuesta. |
| Ficheros de voz | 150 párrafos leídos |
| Referencias | Campione y Véronis (1998); Estruch y Garrido (1996) |

GAUDÍ, "un gran corpus en español para identificación y verificación de hablantes" (Ortega *et al.* 1998). Ha sido desarrollado recientemente en colaboración entre la Escuela Universitaria de Ingenieros de Telecomunicaciones (Universidad Politécnica de Madrid) y el Servicio de Policía Judicial de la Dirección General de la Guardia Civil, con el objetivo específico de contribuir a estudiar la identificación de hablantes.

| | |
|---|---|
| Locutores | 224 hombres -104 de ellos constituyen el subcorpus AHUMADA<br>231 mujeres |
| Canal | - grabación *in situ* (varios micrófonos) en habitación silenciosa<br>- grabación telefónica |
| Tareas | Lectura:<br>- dígitos aislados y concatenados<br>- frases y texto equilibrados (éste con tres ritmos: normal, rápido y lento)<br>- un texto específico para cada sujeto.<br>Habla espontánea:<br>- descripción libre de más de un minuto |
| Ficheros de voz | Sin considerar la lectura de dígitos, 6.825 grabaciones |
| Referencias | Ortega *et al.* 1998; Ortega *et al.* 2000. |

SpeechDat: 4000 locutores, agrupados por edades y modalidad de habla, fueron incorporados a este gran proyecto europeo, que continúa ampliándose en la actualidad. Sin

embargo, su orientación hacia los teleservicios conlleva que el canal de recogida de datos sea exclusivamente telefónico, con las limitaciones acústicas aparejadas.

| | |
|---|---|
| Locutores | 2061 hombres, 1939 mujeres, agrupados por edades (mayoría 15-29) y modalidades de habla (5 regiones de toda España) |
| Canal | telefónico |
| Tareas | Lectura:<br>- dígitos aislados y concatenados, cantidades de dinero<br>- letras (deletreo), palabras aisladas, fechas, nombres propios, frases de tiempo, interrogativas absolutas y 9-10 frases equilibradas o fónicamente ricas<br>Habla espontánea:<br>- un número de teléfono, un día de la semana, una fecha y un sitio<br>Léxico, diccionario fonético y transcripción ortográfica con comentarios |
| Ficheros de voz | 18.036 frases grabadas |
| Referencias | Moreno y Winski (1996). Más información en: `http://www.speechdat.org` |

De las tablas anteriores se deduce la enorme cantidad de datos que se pueden obtener de las cinco bases mencionadas.

Para el análisis del nivel segmental, miles de realizaciones de todos los fonemas del español pueden estudiarse en distintas condiciones de naturalidad, desde estímulos creados "ad hoc", como la lista de pseudo-palabras de EUROM1 (el extremo más artificial, pero también el de mayor control de variables), hasta las muestras de habla espontánea que ofrece GAUDÍ, pasando por el tipo de tarea más frecuente en todas estas bases de datos: la lectura de frases, -fonéticamente equilibradas o fonéticamente ricas- y la lectura de párrafos o textos.

La enorme cantidad de sujetos reclutados para formar la base de SpeechDat permitirá, a pesar de las limitaciones del canal telefónico, analizar factores como la duración segmental.

La influencia de variables como la intensidad o el esfuerzo articulatorio puede analizarse a partir del corpus Lombard en ALBAYZÍN, mientras que los efectos del ritmo de habla o velocidad de elocución son susceptibles de estudio utilizando el texto equilibrado de AHUMADA-GAUDÍ

Sin embargo, el estudio del nivel suprasegmental no ofrece las mismas posibilidades; prácticamente sólo se cuenta con los 20 párrafos con modalidades oracionales interrogativas y exclamativas de EUROM1.

### *3.2 La variación intralocutor*

Para caracterizar los elementos que cambian y los invariables en distintas emisiones de un mismo hablante es preciso grabar a cada sujeto repitiendo los mismos estímulos en diferentes momentos. Así se hizo en EUROM con las pseudo-palabras, que fueron emitidas cinco veces por cada uno de los doce locutores que las leyeron.

En GAUDÍ, tanto las frases como el texto fónicamente equilibrado fueron leídos en tres sesiones de grabación distintas (además de otras tres sesiones realizadas mediante aparato telefónico), controlando el tiempo transcurrido entre ellas (entre 20 y 40 días); en el caso del texto, además pueden analizarse los efectos del ritmo de elocución en la variación del habla de un mismo sujeto, puesto que en cada sesión se les pidió una lectura lenta, otra rápida y otra a ritmo normal.

ALBAYZÍN permite observar qué parámetros resultan modificados y cuáles no cuando 20 locutores leen una serie de frases en silencio y luego las repiten intentando sobreponer su voz a un ruido intenso.

Para el análisis de la variación de estilos intralocutor, es decir, para estudiar qué cambia en cada persona cuando habla espontáneamente o cuando lee, puede recurrirse al minuto de habla espontánea de GAUDÍ, intentando seleccionar fragmentos comparables con los de las diversas tareas de lectura de esos mismos locutores. Sin embargo, ni esta base de datos ni ninguna de las anteriores ha sido diseñada específicamente para comparar la variación inter o intralocutor en distintos estilos de habla, por lo que posiblemente sólo se obtendrán resultados limitados al respecto.

### *3.3 La variación interlocutor*

En este caso, el análisis se enfoca hacia los elementos segmentales y suprasegmentales que difieren de un sujeto a otro, aún emitiendo las mismas secuencias, con el mismo estilo de habla, y sin diferencias geográficas o sociales que puedan ser responsables de la variación; estos serían, en último término, los rasgos que determinan la individualidad de un hablante frente a aquellos que son comunes al habla de un determinado grupo.

Para abordar esta tarea se cuenta con todo el repertorio de locutores de las cinco bases de datos: más de dos mil en SpeechDat, leyendo frases fónicamente ricas, y agrupados por edades y modalidades de habla; 445 en GAUDÍ, en lectura (frases equilibradas y un texto de casi 180 palabras) y una descripción espontánea; 304 en ALBAYZIN, todos hablantes de castellano central, a los que se pidió la lectura de frases equilibradas y 60 de EUROM1 que leyeron frases y párrafos.

En conclusión, el nivel segmental en tareas de lectura está sobradamente representado en los corpus orales disponibles actualmente para el español. El nivel suprasegmental y los estilos de habla más espontáneos requerirían, en una segunda fase del trabajo, la obtención de recursos propios para el proyecto.

## *4. CONCLUSIONES*

La revisión bibliográfica llevada a cabo con el fin de determinar los parámetros acústicos relevantes para el estudio de la variación inter e intralocutor pone de manifiesto la dificultad de encontrar estudios de fonética acústica del español que aborden en profundidad la cuestión del reconocimiento del hablante. Del análisis de la bibliografía disponible en otras lenguas se concluye que existe una cierta unanimidad entre los diversos autores en lo que se refiere a los parámetros acústicos mencionados como responsables de la individualidad de la voz, aunque no parece fácil establecer una prioridad clara entre ellos, puesto que los resultados de cada investigación avalan a uno u otro indistintamente. Con todo, parámetros como el $F_0$, el LTAS y la estructura formántica parecen ser los más influyentes en el reconocimiento de locutor. Se constata, además, que en las introducciones generales a la fonética forense se dedica tanta o más atención al proceso de obtención y tratamiento de las muestras y a sus implicaciones legales que al análisis de los rasgos fonéticos que interesa estudiar en dichas muestras, por lo que no resultan, en principio, especialmente útiles para los objetivos del proyecto.

Teniendo en cuenta la necesidad de reutilización de recursos, se ha llevado también a cabo un análisis de los corpus orales existentes en español valorando su utilidad en relación con los objetivos del proyecto. De esta revisión se desprende que, en general, los datos disponibles para el nivel segmental son ampliamente suficientes para un estudio como el propuesto. La representación, sin embargo, es menor en lo que se refiere a los elementos suprasegmentales, aunque EUROM1 y MULTEXT contienen bastante variación entonativa en el nivel oracional, limitada, sin embargo, a tareas de lectura. Sólo AHUMADA ofrece algo de conversación espontánea, así como muestras de variación temporal, con textos leídos a tres ritmos diferentes. La principal carencia detectada sería un corpus de diálogos espontáneos orales, con suficiente calidad acústica, cuya obtención quedaría para un segundo proyecto, continuación del actual.

***REFERENCIAS***


Atal, B.S. (1972) "Automatic Speaker recognition based on pitch contours", *J.A.S.A* 52, págs. 1687-1697.

Bordone-Sacerdote, C. y Sacerdote, G.G. (1969) "Some spectral properties of individual voices", *Acustica* 21, págs. 199-210.

Brown, R. (1981) "An experimental study of the relative importance of acoustic parameters for auditory speaker recognition", *Language and Speech* 24, 4, págs. 295-310.

Campione, E. y Véronis, J. (1998): "A multilingual prosodic database", *ICSLP'98, Proceedings of the 5th International Conference on Spoken Language Processing*, 30th November-4th December 1998, Sydney, Australia. Vol. 7, págs. 3163-3166.
`http://www.up.univ-mrs.fr/~veronis/pdf/1998icslp-database.pdf`

Carrell, T.D. (1984) "Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification", *Research on Speech Perception Technical Report* (Indiana University Speech Laboratory), 5.

Casacuberta, F., García, R., Llisterri, J., Nadeu, C., Pardo, J.M. y Rubio, A. (1992) "Desarrollo de corpus para investigación en tecnologías del habla (Albayzín)", *Procesamiento del Lenguaje Natural, Boletín nº* 12, págs. 35-42.

Compton, A. J. (1963) "Effects of filtering and vocal duration upon the identification of speakers aurally", *J.A.S.A.* 35, págs. 1748-1752.

Díaz Verdejo, J.E., Rubio, A.J., Segarra, E., Prieto, N. Y Casacuberta F. (1998) "Albayzín: a task-oriented Spanish speech corpus", *Proceedings of the First International Conference on Language Resources and Evaluation*. May 28 - 30, 1998, Granada, España. European Language Resources Association. Vol. I, págs. 497-502.

Doddington, G.R. (1985) "Speaker recognition. Identifying people by their voices", *Proc. IEEE* 73, págs. 1651-1664.

Doherty, E. T. (1976) "An evaluation of selected acoustic parameters for use in speaker identification", *Journal of Phonetics* 4, págs. 321-326.

Doherty, E. y Hollien, H (1978) "Multiple factor speaker identification of normal and distorted speech", *Journal of Phonetics* 6, págs. 1-8.

Dommelen, W. A. van (1997) "The contribution of speech rhythm and pitch to speaker recognition", *Language and Speech* 30, 4, págs. 325-338.

Eskenazi, M., Childers, D.G. y Hicks, D.M. (1990) "Acoustic correlates of vocal quality", *Journal of Speech and Hearing Research* 33, págs. 298-306.

Estruch, M. y Garrido, J.M. (1996) "Report on Prosody Tools Efficiency and Failures (Spanish EUROM)", Prosody Tools Efficiency and Failures. WP 4 Corpus. T4.6 Speech Markup and Validation. Deliverable 4.5.2. Final version. 15 October 1996. LRE Project 62-050 MULTEXT.
`http //liceu.uab.es/~joaquim/publicacions/Prosody_tools_96.pdf`

Furui, S. (1978) "Effects of Long-Term Spectral Variability in Speaker Recognition", *J.A.S.A.* 64: S183.

Furui, S. (1986) "Research on individuality features in speech waves and automatic speaker recognition techniques", *Speech Communication* 5, 2, págs. 183-197.

Gelfer, M.P. *et alii* (1989) "The effects of sample duration and timing on speaker identification. Accuracy by means of long-term spectra", *Journal of Phonetics* 17:4, págs. 327-338.

Glenn, J. W. y Kleiner, N. (1968) "Speaker identification based on nasal phonation", *J.A.S.A.* 43, págs. 368-372.

Gobl, C. (1989) "A preliminary study of acoustic voice quality correlates", *STL-Quarterly Progress Status Report* 4, págs. 9-22.

Hollien, H. (1990) *The Acoustics of Crime. The New Science of Forensic Phonetics*, Plenum, Nueva York.

Hollien, H. (1991) "The profile approach to speaker identification", *Actes du XII$^{ème}$ Congrès International des Sciences Phonétiques* (Aix-en-Provence, 1991), Université de Provence, Aix, págs. 396-401.

Hollien, H. y Majewski , W. (1977) "Speaker identification by long-term spectra under normal and distorted speech", *J.A.S.A.* 62, págs. 975-980.

Hollien, H., W. Majewski y E.T. Doherty (1982) "Perceptual identification of voice under normal, stress and disguise speaking conditions", *Journal of Phonetics* 10, págs. 139-148.

Ingram, J.C.L., Prandolini, R. y Ong, S. (1996) "Formant trajectories as indices of phonetic variation for speaker identification", *Forensic Linguistics*, vol. 3-1, 129-145.

Itoh, K. y Saito, S. (1982) "Effects of acoustical feature parameters of speech on perceptual identification of speaker", *IECE Trans*. Vol. J65-A, págs. 101-108.

Johnson, C.C., Hollien, H. y Hicks, J.W. Jr. (1984) "Speaker identification utilizing selected temporal speech features", *Journal of Phonetics* 12, págs. 319-327.

Karlsson, I. (1988) "Glottal waveform parameters for different speaker types", *Proc. Speech '88, 7$^{th}$ FASE Symposium*, vol. 1, págs. 225-231.

Kreiman, J. y Papcun, G. (1991) "Comparing discrimination and recognition of unifamiliar voices", *Speech Communication* 10, págs. 265-275.

Künzel, H. J. (1995) "Field procedures in forensic speaker recognition", en J. Windsor Lewis (Ed.) *Studies in General and English Phonetics*, Routledge, Londres, págs. 68-84.

Kuwabara, H. y Takagi, T. (1991) "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method", *Speech Communication* 10, págs. 491-495.

Kuwabara, H. y Ohgushi, K. (1987) "Contributions of vocal tract resonants frequencies and bandwidths to the personal perception of speech", *Acustica* 63, págs 121-128.

Kuwabara, H. y Sagisaka, Y. (1995) "Acoustic characteristics of speaker individuality: Control and conversion", *Speech Communication* 16, págs. 165-173.

Ladefoged, P. y Broadbent, D.E. (1957) "Information conveyed by vowels", *J.A.S.A.* 29, págs. 98-104.

LaRiviere, C. (1975) "Contribution of fundamental frequency and formant frequencies to speaker identification", *Phonetica* 31, págs. 185-197.

Llisterri, J., Aguilar, L., Blecua, B., Machuca, M.J., de la Mota, C., Ríos, A., Moreno, A. y Salavedra, J. (1993) *Spanish EUROM 1: Phonetic Contents*. Report D6 Appendix X. SAM-A/UPC/002. ESPRIT PROJECT 6819 (SAM-A) Speech Technology Assessment in Multilingual Applications.

Matsumoto, H. *et alii* (1973) "Multidimensional representation of personal quality of vowels and its acoustical correlates", *IEEE Trans.* Vol. AU, 21, págs. 428-436.

Miller, J.E. (1964) "Decapitation and recapitation: a study of voice quality", *J.A.S.A.* 36.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. y Nadeu, C. (1993) "ALBAYZÍN Speech Database: Design of the Phonetic Corpus", in *Eurospeech'93. 3rd European Conference on Speech Communication and Technology.* Berlin, Alemania, 21-23 September 1993. Vol. 1, págs. 175-178.

Moreno, A. y Winsky, R. (1996) SpeechDat (M) Spanish Database. *SpeechDat CD-ROM.* ELDA, European Language Resources Distribution Agency.

Ortega, J., González, J. y Marrero, V. (2000) "AHUMADA: A large corpus in Spanish for speaker characterization and identification", *Speech Communication* 31, 2-3: 255-264.

Ortega, J., González, J, Marrero, V., Díaz, J.J., García, R., Lucena, J. y Sánchez, J.A.G. (1998) "AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification", *Proceedings of ICAPSSP-98. IEEE International Conference on Acoustics Speech and Signal Processing.* May 1998. págs. 773-776.
http://www.atvs.diac.upm.es/publicaciones/docs/Ort98b.pdf

Pisoni, D. B. (1993) "Long-term memory in speech perception: some new findings on talker variability, speaking rate and perceptual learning", *Speech Communication* 13, págs. 109-125.

Pittam, J. (1987) "The long-term spectral measurement of voice quality as a social and personality marker: a review", *Language and Speech* 30, págs. 1-13.

Pollack, I., J.M. Pickett y W.H. Sumby (1954) "On the identification of speakers by voice", *J.A.S.A.* 26, págs. 403-412.

Pruzansky, S. (1963) "Pattern matching procedure for automatic for automatic talker recognition", *J.A.S.A.* 35, págs. 354-358.

Ramón, J. L., Garcerán, V., Canteras, M. y Sánchez Molero, J.A. (2000) "Parametric speaker verification with linear prediction and cepstrum using the envelope of voice and discriminant analysis", *Actas del I Congreso de la Sociedad Española de Acústica Forense*, 5 y 6 de octubre de 2000, págs. 169-181.

Shearme, J.N. y J.N. Holmes (1959) "An experiment concerning the recognition of voices", *Language and Speech* 2, págs. 123-131.

Stevens, K. (1971) "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds", *Proceedings of the 7th International Congress of. Phonetic Sciences*, Montreal, La Haya, Mouton, págs. 206-227.

Stevens, K.N., Williams, C.E., Carbonell, J.R. y Woods, B. (1968) " Speaker identification and authentication: a comparison of spectrographic and auditory presentation of speech materials", *J.A.S.A.* 44, págs. 1596-1607.

Stevens, K.N. y House, A.S. (1963) "Perturbations of Vowel Articulations by Consonantal Context: An Acoustical Study", *Journal of Speech and Hearing Research* 6,2, págs. 111-128.

Su, L. S, Li, K. P. y Fu, K. S. (1974) "Identification of speakers by use of nasal coarticulation", *J.A.S.A.* 56, págs. 1867-1882.

Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C., Nichol, J. y Nash, W. (1972) "Experiment on voice identification", *J.A.S.A.* 51, págs. 2030-2043.

Wolf, J. J. (1972) “Efficient acoustic parameters for speaker recognition”, *J.A.S.A.* 51, págs. 2044-2056.