Base de Datos Audiovisual y Multicanal en Castellano para Reconocimiento Automático del Habla Multimodal en el Automóvil.

Alfonso Ortega[1], *Federico Sukno*[1], *Eduardo Lleida*[1],
Alejandro Frangi[2], *Antonio Miguel*[1], *Luis Buera*[1],*Ernesto Zacur*[2]

[1]Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza.
[2] Departamento de Tecnología, Universidad Pompeu Fabra.
{ortega,fsukno,lleida,afrangi,amiguel,lbuera,zacur}@unizar.es

Resumen

En este trabajo se describe la adquisición de la base de datos audiovisual y multicanal en castellano AV@CAR. El propósito de este corpus es servir de material para el estudio y desarrollo de sistemas de reconocimiento automático del habla multimodal en el entorno del automóvil. Se trata éste de un entorno donde el reconocimiento automático del habla juega un papel muy importante debido a que el uso de estas tecnologías puede evitar la distracción del conductor en gran número de ocasiones.

El interior del coche es un ambiente ruidoso donde el apoyo de información visual (lectura de labios) puede elevar las prestaciones de los sistemas de reconocimiento automático del habla.

El corpus multimodal está compuesto por siete canales de audio que incluyen, voz limpia (capturada mediante un micrófono de cercanía, close-talk), voz de micrófonos distantes situados en el techo del habitáculo, ruido de referencia proveniente del motor e información acerca de la velocidad del vehículo.

La parte visual de la base de datos está capturada usando una pequeña cámara de vídeo de bajo coste situada en el parabrisas del vehículo, junto al espejo retrovisor del mismo. Las grabaciones se realizan bajo diferentes situaciones de conducción e iluminación: coche parado, tráfico urbano, autovía, pavimento en mal estado, conducción nocturna y conducción de día.

El corpus audiovisual que se presenta contiene también una parte del mismo que ha sido adquirido en el laboratorio bajo condiciones acústicas y de iluminación controladas.

1. Introducción

La seguridad en la conducción es uno de los aspectos más importantes a tener en cuenta a la hora de diseñar aplicaciones destinadas a ser utilizadas por el conductor de un vehículo. Por este motivo, la introducción de las tecnologías del habla, síntesis de voz y reconocimiento automático del habla, se impone a la hora de desarrollar los interfaces de usuario para dispositivos y servicios que centran su ámbito de aplicación en el automóvil (teléfonos móviles, asistentes digitales personales, sistemas de navegación, etc.).

Sin embargo, el uso de este tipo de interfaces dentro de los coches supone un reto muy importante. El alto nivel de ruido presente en el interior del habitáculo y la distancia entre el locutor y los sensores encargados de captar su voz degradan en gran medida las prestaciones de los sistemas de reconocimiento.

Figura 1: *Grabación del corpus del vehículo*

Como apoyo al reconocedor, puede incluirse otro tipo de información no acústica que eleve las tasas de acierto del mismo. Las técnicas de lectura de labios pueden servir de ayuda para este tipo de sistemas en ambientes altamente ruidosos. Asimismo, contar con señales provenientes de más de un sensor permiten la utilización de algoritmos que eleven la calidad de la entrada o entradas al reconocedor. Puede encontrarse información actualizada al respecto de las prestaciones de los sistemas de reconocimiento automático del habla bimodales en [1] y [2]

Para el desarrollo de este tipo de aplicaciones, es imprescindible contar con bases de datos que combinen la información tanto de vídeo como de audio de forma síncrona. Existen varias bases de datos de este tipo, la mayoría de ellas en lengua inglesa [3, 4] aunque pueden encontrarse ejemplos de bases de datos bimodales en otros idiomas como el holandés [5] o la base de datos BANCA [6] que incluye adquisiciones en castellano, francés, inglés e italiano. La diferencia fundamental entre la mayoría de estos córpora y el que se presenta en este trabajo, es el ámbito de grabación de los mismos. Mientras que las bases de datos anteriormente descritas, han sido adquiridas en un ambiente de laboratorio, la presente base de datos está compuesta por una parte grabada en condiciones de laboratorio y otra que se adquiere en un vehículo a motor en condiciones nor-

Este trabajo ha sido parcialmente financiado por los proyectos TIC2002-04495-C02 y TIC2002-04103-C03-01 del Ministerio de Educación y Ciencia y por Vision RT Ltd.(GB). Alejandro Frangi y Luis Buera reciben financiación a través de becas personales, Ramón y Cajal y FPU, respectivamente, del Ministerio de educación y Federico Sukno del BSCH y la Universidad de Zaragoza.

males de conducción. No se tiene conocimiento de la existencia de ninguna otra base de datos bimodal grabada en un coche en castellano. Sí existe una base de datos adquirida dentro de un vehículo en checo [7] pero se trata de un corpus monocanal, adquirido con una cámara de video digital de alta calidad y alto coste.

Hasta el momento, el trabajo de investigación en el ámbito del reconocimiento automático del habla audiovisual se ha centrado en el empleo y medida de prestaciones sobre bases de datos adquiridas bajo condiciones visuales ideales. Estas bases de datos cuentan con video de alta resolución, imagen frontal, variaciones limitadas en la posición y el gesto de la cara del locutor, distancia casi constante entre la cámara y el sujeto, iluminación prácticamente uniforme y en la mayoría de los casos, fondo uniforme. A diferencia de estas idealidades en la parte visual de la base de datos, el canal de audio era degradado artificialmente con ruido aditivo. Para verificar los beneficios que el empleo de información visual aporta al reconocimiento automático del habla es necesario medir las prestaciones de los sistemas audiovisuales en entornos que carezcan de esas idealidades en el canal visual [8]. Con este objetivo se pretende disponer de la suficiente cantidad de datos audiovisuales en escenario real como para llevar a cabo este tipo de tareas. Los datos recogidos en el automóvil se caracterizan por variaciones de la posición de la cara del locutor, de iluminación y del fondo; por equipamiento de adquisición de bajo coste o por la presencia de sombras cambiantes sobre el rostro del locutor.

El material proporcionado por la presente base de datos permite no sólo el estudio y desarrollo de sistemas de reconocimiento automático del habla audiovisual, sino que también al poseer señales provenientes de varios sensores y de varias fuentes (voz y ruido), permite el estudio de técnicas y algoritmos de reducción de ruido, adaptación al entorno, compensación de modelos, etc. encaminadas a conseguir sistemas más robustos ante situaciones de alto nivel de ruido ambiente.

Este trabajo presenta la siguiente organización: En la sección 2 se describe la base de datos AV@CAR de acuerdo a las diferentes partes que la componen, grabaciones en el laboratorio, grabaciones en el vehículo, el instrumental empleado para la adquisición del audio y el equipamiento utilizado para la captura de la parte de video. En la sección 3 se describen los procedimientos seguidos durante la adquisición de la misma así como las tareas que la componen. Finalmente, en la sección 4 se presentan las conclusiones.

2. Descripción de la base de datos

La base de datos audiovisual AV@CAR, puede dividirse en dos partes fundamentales. La primera de ellas está grabada en un coche en condiciones normales de conducción y la segunda se adquiere en un entorno libre de ruido y con condiciones de iluminación controladas en un ambiente de laboratorio.

La parte de la base de datos grabada en el vehículo está compuesta por siete canales de audio, un canal de vídeo e información acerca de la velocidad del vehículo en todo momento. Asimismo, se incluye información de las condiciones de la carretera, la climatología, el locutor y la iluminación. Esto permite la definición de diferentes entornos o escenarios en base a los cuales estudiar la aplicación de técnicas y algoritmos de adaptación al locutor y/o adaptación al entorno [9].

Por otro lado, la base de datos consta de una segunda parte adquirida en el laboratorio donde se han recogido las señales de cinco micrófonos, una señal de video e imágenes tridimensionales de cada locutor para propósitos biométricos.

Figura 2: *Posición de los micrófonos situados sobre los asientos delanteros.*

Figura 3: *Posición de los micrófonos situados sobre los asientos traseros.*

2.1. Corpus de audio adquirido en el vehículo

Para adquirir siete canales de forma síncrona junto con la información de velocidad del vehículo se ha empleado el sistema de adquisición de ocho canales de 24 bits *Hammerfall DSP Multiface* de RME (Alemania). Este sistema permite la rápida transferencia de datos con un ordenador personal de sobremesa a través de una tarjeta con interfaz PCI o con un ordenador portátil a través de una tarjeta con interfaz PCMCIA. Durante la grabación se ha empleado un equipo de acondicionamiento de señal de ocho canales *octamic* también de RME (Alemania). Las grabaciones han sido realizadas empleando una fuente de corriente continua de 12 V convenientemente aislada de la batería del vehículo para evitar los ruidos e interferencias provenientes del sistema eléctrico del mismo.

Los micrófonos elegidos han sido los Q501T (AKG, Austria) debido a su respuesta frecuencial paso alto que los hacen adecuados para su empleo en el interior del automóvil. Se han instalado seis micrófonos en el interior del coche, situándolos en el techo del mismo, tres en la parte delanteras (dos en la posición del conductor) y tres en la parte trasera. En las figuras 2 y 3 pueden observarse las ubicaciones de los mismos.

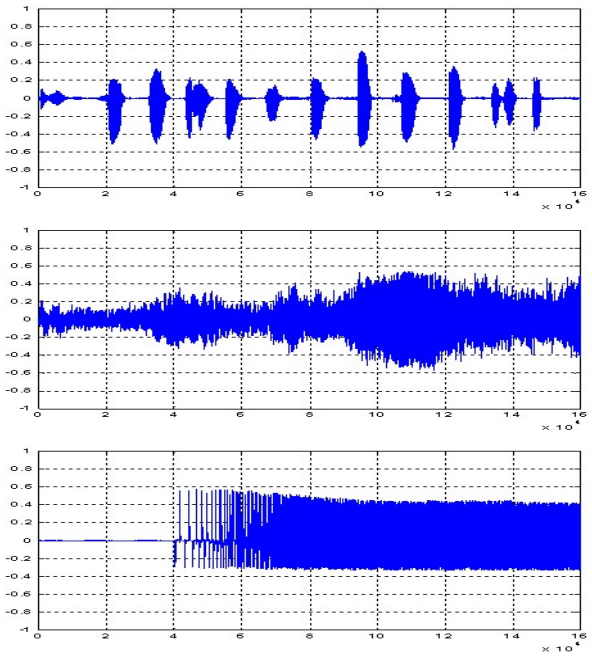

Figura 4: *Señales de ejemplo adquiridas en el vehículo. Señal de Close-Talk (superior). Señal del micrófono situado sobre el conductor (central). Señal PWM que contiene la información de velocidad (inferior).*

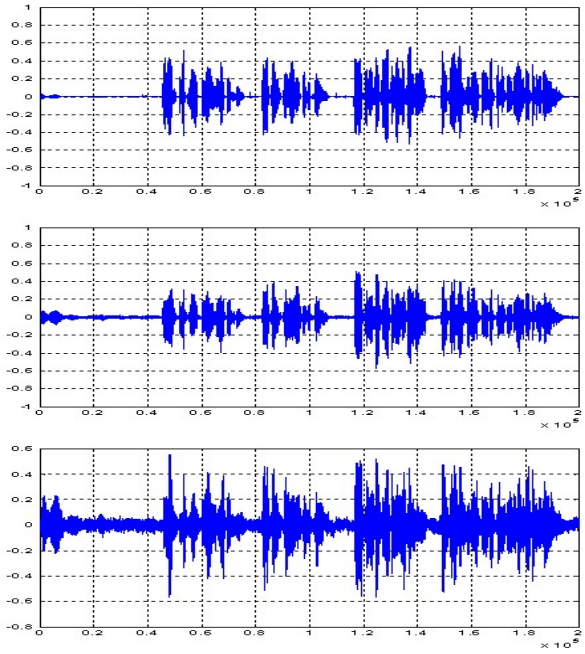

Figura 5: *Señales de ejemplo adquiridas en el laboratorio. Señal de Close-Talk (superior). Señal del micrófono situado a 30 cm del locutor (central). Señal de uno de los micrófonos situados en las esquinas superiores de la sala (inferior).*

Una señal de voz de alta relación señal a ruido se captura empleando un micrófono de cercanía (close-talk) C444L (AKG, Austria).

Uno de los canales de entrada del sistema de adquisición se reserva para capturar la señal proveniente de un micrófono instalado en el compartimiento del motor con el fin de contar con una señal de referencia sólo de ruido y poder estudiar y desarrollar algoritmos de cancelación de ruido con referencia u otro tipo de técnicas multicanal que requieran de una señal que contenga ruido pero no señal de voz.

La información de la velocidad del vehículo se adquiere de forma síncrona con el resto de los canales de voz y ruido gracias a la adquisición de una señal PWM cuyo periodo es proporcional a la velocidad del coche.

Pueden observarse ejemplos de las señales adquiridas en el coche en la figura 4

2.2. Corpus de audio adquirido en el laboratorio

Para las sesiones adquiridas en el laboratorio se han empleado diferentes tipos de micrófonos. Por un lado se ha empleado un micrófono igual a los empleados en el vehículo Q501T (AKG, Austria) situado a unos 30 cm del locutor. La señal de voz limpia se adquiere con un micrófono close-talk C 477 W R (AKG, Austria).

Para la captura con sensores en campo lejano se han elegido los micrófonos C 417 (AKG, Austria) y CK 80 (AKG, Austria) situados en las esquinas superiores de la sala de grabación. Esta sala tiene unas dimensiones de $1{,}86 \times 2{,}86 \times 2{,}11\ m$

Durante las sesiones de grabación en el laboratorio se han adquirido señales empleando una cabeza y un torso 4100 D (Brüel&Kjaer, Dinamarca) con los amplificadores de acondicionamiento *Nexus 2693* para completar una base de datos binaural.

Toda la base de datos de audio está muestreada a 16 kHz y almacenada con 16 bits cada una de las muestras.

En la figura 5 se muestran ejemplos de las señales adquiridas en el laboratorio.

2.3. Corpus de video adquirido en el vehículo

Para la parte de video grabada en el coche, se ha optado por usar una pequeña cámara de bajo coste V-1204A (Marshall Electronics, USA) sensible a las longitudes de onda del espectro visible y a las del infrarrojo cercano. Esto permite su utilización incluso en situaciones de muy baja iluminación como puede ser

Figura 6: *Posición de la cámara junto al espejo retrovisor.*

Figura 7: *Ejemplo de un frame adquirido en el vehículo*

la conducción interurbana nocturna.

La videocámara está situada en el parabrisas del vehículo junto al espejo retrovisor para no reducir el campo de visión del conductor y a la vez capturar en todo momento la cara del mismo.

La cámara incluye seis LEDS que emiten en el infrarrojo y que proporcionan la suficiente iluminación como para poder usar la cámara sin necesidad de contar con ningún otro tipo de fuente de iluminación adicional.

Las imágenes tomadas son de 8 bits en escala de grises. La resolución espacial es de 768×576 pixels con una tasa de refresco de 25 imágenes por segundo. Estas imágenes son digitalizadas empleando una tarjeta *DT3120 Frame Graber* (Data Translation Inc., USA). La alta resolución espacial está justificada teniendo en cuenta que la imagen debe contener la cabeza del conductor en todo momento y por lo tanto debe contemplar los movimiento de este ante cualquier situación que se presente durante la conducción.

2.4. Corpus de video adquirido en el laboratorio

Las grabaciones realizadas en el laboratorio están divididas en dos partes. En la primera, el locutor es grabado cuando pronuncia alguna de las palabras o frases de alguna de las tareas del corpus de audio. La adquisición de esta parte del corpus se realiza utilizando el mismo modelo de cámara que se emplea para las grabaciones en el vehículo (V-1204A) pero en esta ocasión la iluminación está controlada en todo momento y la toma de la cara del locutor es completamente frontal a diferencia de la toma semilateral que se obtiene con la cámara del vehículo situada en la esquina superior derecha.

En la segunda parte del corpus de video del laboratorio, se toman varias imágenes del locutor mediante un equipo de captura tridimensional de Vision RT Ltd (London, UK), compuesto por dos conjuntos de tres cámaras cada uno. Este equipo de adquisición toma simultáneamente seis imágenes de cada individuo. Con cuatro de ellas se realiza una reconstrucción 3D de la geometría facial y con las dos restantes se obtiene información de la textura de la cara en blanco y negro. De esta manera, se obtienen superficies tridimensionales con textura de la cara de cada locutor.

Cada uno de los individuos que participaron en la fase de adquisición posaron con diferentes expresiones faciales basadas en la clasificación gestual de Ekman [10] y Martínez [11] y

Figura 8: *Grabación del corpus del laboratorio*

fueron fotografiados desde distintos ángulos. Esta información sera útil al comparar las imágenes obtenidas en el vehículo (visión semilateral) con las frontales obtenidas en el laboratorio y presentes en la mayoría de las bases de datos faciales.

La base de datos tridimensional también se grabó con una cámara de video de color 1352-5000 (Cohu Inc. USA) provista de una lente de aumento Navitar TV (12.5-75 mm, F 1.8). La digitalización de estas imágenes se ha realizado con 768×576 pixels, 24 bits y 25 frames por segundo.

2.5. Sincronización del video con el audio.

Uno de los retos a la hora del diseño y la obtención de esta base de datos fue el asegurar un sincronismo suficientemente preciso entre la parte de audio de la misma y la parte de video.

Para conseguir la suficiente independencia de retardos no controlables del hardware o de los sistemas operativos de los ordenadores empleados durante la adquisición, se colocó junto al locutor una tira de LEDs. Esta tira de LEDs, está presente en una esquina del encuadre en cada uno de los frames de video que componen la base de datos como puede observarse en las figuras 7 y 10.

Estos LEDs se encienden y se apagan secuencialmente cada 5 ms de manera síncrona con una señal eléctrica generada por el sistema de entrada y salida de audio. Asimismo, la parte de audio del sistema de adquisición produce un estímulo acústico a través de los altavoces del vehículo o del laboratorio al comienzo de cada grabación de forma simultánea con la señal eléctrica que enciende los LEDs. Posteriormente, puede realizarse un alineamiento preciso de la información proveniente de las ambas fuentes gracias a la sincronía entre el estímulo acústico recogido

Figura 9: *Equipo de adquisición 3D*

Figura 10: *Ejemplo de un frame adquirido en el laboratorio*

por los micrófonos del equipo de adquisición y la iluminación de los LEDs capturados en cada uno de los frames de la parte de video del corpus.

Los LEDs parpadean durante toda la grabación cada segundo, a intervalos controlados por el sistema de entrada/salida de audio para permitir la verificación del sincronismo.

3. Procedimientos y tareas de la grabación

El presente corpus audiovisual puede dividirse en tres grupos principales atendiendo a sus tareas y entornos. Así, se dispone de una parte destinada principalmente a tareas de entrenamiento y adaptación adquirida en el vehículo. Otra parte formada por datos dependientes de la aplicación destinada para labores de test también adquirida en el vehículo. Y una tercera grabada en el laboratorio, un entorno libre de ruido y con iluminación controlada. Cada una de estas partes se compone de diversas tareas.

La parte de la base de datos grabada en el coche para entrenamiento y adaptación se compone de cuatro tareas:

1. Lectura de un texto largo con el vehículo estacionado y el motor apagado.
2. Repetición de 25 frases fonéticamente balanceadas con el vehículo estacionado y el motor apagado.
3. Repetición de 25 frases fonéticamente balanceadas en condiciones de conducción.
4. Grabación de señales de ruido bajo diferentes circunstancias de conducción y estados de tráfico.

La segunda parte de la base de datos dependiente de la aplicación, también está compuesta de cuatro tareas:

1. Repetición de frases y palabras específicas de la aplicación (principalmente relacionadas con el uso de un terminal celular, un sistema de navegación para vehículo o acceso remoto a servicios como el correo electrónico)
2. Deletreo
3. Dígitos y números.
4. Nombres de calles, ciudades o regiones.

Las principales características del corpus audiovisual adquirido en el vehículo son:

1. Entorno acústico altamente ruidoso y muy cambiante. La relación señal a ruido de la voz captada por los micrófonos en el coche, es muy variante y depende tanto del locutor como de las características del vehículo o de la vía en la cual se está circulando. Asimismo, tanto la distribución espectral del ruido como su potencia dependen de muchas circunstancias como pueden ser el tipo de pavimento sobre el que se rueda, la velocidad de circulación, el empleo de climatizadores o aires acondicionados, situación de las ventanillas del vehículo, etc.
2. Condiciones visuales reales. Esto implica que la posición de la cámara queda restringida a ubicaciones en las que no limite el campo visual del conductor y por tanto queda descartado el enfoque frontal. Otras características son que el fondo de la imagen sea cambiante, que la iluminación al no ser controlada pueda hacer aparecer sombras no estáticas sobre el rostro del locutor, que las condiciones de luz sean muy dispares cubriendo un amplio espectro que va desde iluminación frontal diurna hasta la conducción nocturna por vías interurbanas no iluminadas.

La tercera parte de la base de datos se adquiere en condiciones de laboratorio contando con los mismos individuos que forman parte del corpus adquirido en el vehículo. Algunas de las tareas realizadas en el vehículo se repiten en el entorno controlado del laboratorio.

El objetivo fundamental de esta parte de la base de datos es el disponer de datos sin el ruido ni las variaciones de iluminación que existen en el coche. Además, en el laboratorio no existen restricciones acerca de la ubicación de la cámara de video y por tanto, es factible el disponer de una imagen frontal del locutor. Los resultados obtenidos con este corpus podrán compararse con los de otras bases de datos estándar adquiridas también en el laboratorio así como estudiar las diferencias con las otras dos partes de la base de datos grabadas en el vehículo.

Además de la parte de datos audiovisual adquirida en el laboratorio, se dispone de imágenes tridimensionales con textura y de video en color de cada locutor con diferentes expresiones faciales tal y como se recoge en la tabla 1

Para las dos primeras partes de la base de datos adquiridas en el vehículo, los procedimientos seguidos durante la adquisición son muy importantes. El texto de las tareas que deben ser realizadas mientras se conduce no puede ser leído por el locutor por motivos de seguridad. Además, la legislación española

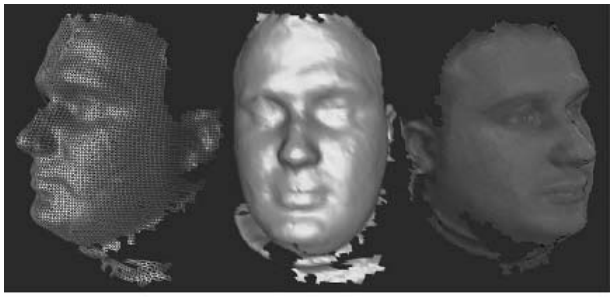

Figura 11: *Imagen facial tridimensional con textura (laterales) y sin textura (frontal)*

Tabla 1: *Tabla de expresiones faciales.*

a) Vista Frontal	h) Felicidad
b) Perfil Izquierdo	i) Sorpresa
c) Perfil Derecho	j) Bostezo
d) Vista Superior	k) Enfado
e) Vista Inferior	l) Disgusto
f) Imagen con gafas transparentes	m) Miedo
g) Imagen con gafas de sol	n) Pena

prohibe durante la conducción el uso de pantallas que puedan ser fuente de distracción para el conductor. Por este motivo, cada frase o palabra que el locutor debe pronunciar es leída en voz alta por el técnico de grabación seguida de un estímulo acústico que indica el inicio de la grabación.

Con el objetivo de estudiar y desarrollar algoritmos de adaptación al entorno y adaptación al locutor, se consideró la conveniencia de contar con una gran cantidad de datos de un número reducido de locutores en diferentes situaciones y entornos. La base de datos consta de 20 personas, 11 hombres y 9 mujeres, cuyas edades varían entre los 25 y los 50 años. En cuanto a las condiciones de iluminación en la parte de los datos adquirida en el vehículo, aproximadamente la mitad de las sesiones de los locutores varones son diurnas y la otra mitad nocturnas. De igual manera se distribuyeron las sesiones realizadas con locutores de sexo femenino.

4. Conclusiones

En este trabajo se ha presentado la base de datos audiovisual, multicanal AV@CAR. El objetivo fundamental de este corpus es la dotación de material útil para el estudio y diseño de sistemas de reconocimiento automático del habla audiovisuales en el entorno del automóvil.

Debido al alto nivel de ruido presente en el interior del habitáculo, la información visual puede elevar las prestaciones del sistema de reconocimiento automático del habla mediante la aplicación de técnicas de lectura de labios (lip-reading).

Adicionalmente, este corpus cuenta con una parte de sus datos adquirida en condiciones de laboratorio con el objetivo de disponer de señales libres de ruido o variaciones de iluminación. En esta parte adquirida en el laboratorio se han incluido imágenes tridimensionales con textura de la cara de cada locutor.

5. Referencias

[1] Potamianos, G., Neti, C., Gravier, G., Garg, A. and Senior, A.W., Recent Advances in the Automatic Recognition of Audiovisual Speech, Proceedings of the IEEE, Vol. 91, No. 9, September 2003, pp. 1306-1326.

[2] Potamianos, G. Neti, C., Luettin, J. and Mathews, I., Audio-Visual Automatic Speech Recognition: An Overview. In Issues in Visual and Audio-Visual Speech Processing, Bailly,G., Vatikiotis,E. and Perrier, P. (Eds), MIT Press (In Press) 2004.

[3] Potamianos, G. Cosatto, E. Graf, H.P. and Roe, D.B. “Speaker Independent Audio-Visual Database for Bimodal ASR, in Proceedings of Eurospeech 2001 (CD-ROM), Aalborg, Denmark 2001.

[4] Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., XM2VTSDB: The Extended M2VTS Database, in 2nd International Conference AVBPA, Washington D.C. 1999.

[5] Wojdel, J.C., Wigges, P., Rothkrantz, L.J.M., An AudioVisual Corpus for Multimodal Speech Recognition in Dutch Language, in Proceedings of ICSLP 2002 (CD-ROM), Denver USA, 2002.

[6] Bailly-Baillire, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., Thiran, J.P., The BANCA Database and Evaluation Protocol, in 4th International Conference AVBPA, Springer-Verlag, 2003.

[7] Zelezný, M y Císar, P., Czech Audio-Visual Speech Corpus of a Car Driver for In-Vehicle Audio-Visual Specch Recognition, in Proceedings of VP 2003, St. Jorioz, France 2003.

[8] Potamianos, G. y Chalapathy, N., Audio-Visual Speech Recognition in Challenging Environments, in Proceedings of Eurospeech 2003 (CD-ROM), Geneve, Switzerland.

[9] Buera, L., Lleida, E., Miguel, A. y Ortega, A., Multi-Environment Model Based Linear Normalization for Speech Recognition, in Proceedings of ICASSP 2004 (CD-ROM), Montreal, Canada, May 2004.

[10] Ekman, P., Friesen, W., Understanding the Face, A Guide to Recognising Emotions from Facial Expressions, Prentice-Hall, 1975.

[11] Martínez, A., Benavente, R., The AR Face Database. Technical Report, Computer Vision Center, Barcelona, Spain, 1998.