Análisis y Síntesis Audiovisual para Interfaces Multimodales Ordenador-Persona

Xavier Sevillano, Javier Melenchón, Joan C. Socoró

Departamento de Comunicaciones y Teoría de la Señal
Enginyeria La Salle – Universitat Ramon Llull
Pg. Bonanova 8 – 08022 Barcelona
{xavis, jmelen, jclaudi}@salle.url.edu

Abstract. Las técnicas multimodales de procesamiento de señal están llamadas a jugar un papel preponderante en la implementación de interfaces que faciliten la interacción natural entre los ordenadores y los humanos. En particular, el desarrollo de interfaces eficientes que emulen la comunicación entre personas requiere el uso de técnicas capaces de procesar conjuntamente los modos visual y auditivo. Este trabajo introduce el uso de herramientas de análisis audiovisual basadas en el análisis en componentes principales y en descomposiciones que utilizan restricciones de no negatividad sobre secuencias faciales de vídeo con voz. Además, se estudia y compara la aplicabilidad de las bases audiovisuales obtenidas mediante estas técnicas a lo largo de diversos experimentos en los que se evalúa la calidad de la síntesis audiovisual de forma objetiva y subjetiva.

1 Introducción

Las interfaces de comunicación actuales entre personas y ordenadores están lejos de asemejarse a la interacción natural entre humanos. Esta situación se traduce en una barrera para un amplio sector de la población en el acceso al creciente volumen de información digital, sea por su poca familiaridad con las tecnologías de la información o por sufrir alguna incapacidad funcional, dando pie a la llamada brecha digital. Por ello, el desarrollo de interfaces sencillas e intuitivas de utilizar se contempla como una posibilidad para reducir esta brecha. En este contexto, las interfaces audiovisuales constituyen una opción factible para permitir una interacción persona-ordenador (IPO) natural, al emular el lenguaje natural propio de la comunicación interpersonal. El diseño de estas interfaces debe considerar los flujos audiovisuales de entrada (persona a ordenador) y de salida (ordenador a persona).

Los trabajos previos realizados por nuestro grupo de investigación en el ámbito de la IPO se han centrado en el desarrollo de interfaces ordenador-persona basadas en cabezas parlantes realistas con animación sincronizada a través de una fuente de texto o voz. Hasta el momento, la voz y la imagen se han sintetizado de forma separada, requiriendo de un proceso de sincronización entre ambos [11]. No obstante, las limitaciones de esta metodología han resaltado la necesidad de llevar a cabo una integración más eficiente basada en la fusión a priori de los modos acústico y visual.

En los últimos años, las técnicas de fusión audiovisual (AV) han abierto un prometedor campo para la investigación. La mayoría de las aproximaciones propuestas en este ámbito analizan las dependencias entre los modos auditivo y visual de secuencias AV para, por ejemplo, localizar las regiones del vídeo donde se origina el sonido. El concepto subyacente en estas aproximaciones es la maximización de la información mutua entre ambos modos y ha sido tratado mediante distintas técnicas, tales como estimadores de la covarianza AV [5], Análisis de Correlación Canónica [13], estimadores no parámetricos de densidades [3] o cadenas de Markov [1].

Por otro lado, técnicas de representación de datos basadas en restricciones de independencia estadística han sido aplicadas con éxito a la detección de correlaciones AV operando sobre un espacio de datos audiovisuales [14].

Este trabajo presenta la aplicación novedosa de dos técnicas al análisis audiovisual conjunto de una secuencia de vídeo: el Análisis en Componentes Principales [7] y la Factorización de Matrices No Negativas [9], descritas en la sección 2. Como resultado del proceso de análisis audiovisual, se obtiene un conjunto de bases AV e información temporal relativa a su activación. Como se muestra en la sección 3, estos datos pueden ser empleados para realizar síntesis AV de un modo relativamente sencillo. La viabilidad de ambas propuestas se estudia en la sección 4 a través de un conjunto de experimentos comparativos iniciales de análisis y síntesis AV. Finalmente, las conclusiones de este trabajo se presentan en la sección 5.

2 Extracción de bases audiovisuales

Un conjunto multidimensional de datos (denotado genéricamente en adelante mediante una matriz **X**) puede ser representado mediante diversas técnicas de transformación de atributos. No obstante, la utilidad de dichos métodos de representación depende tanto de su idoneidad respecto a la naturaleza de los datos como de la utilización que se quiera hacer de los mismos. Por ejemplo, el Análisis en Componentes Principales (*Principal Component Analysis* ó PCA) es útil, por ejemplo, para describir los datos sobre un espacio ortogonal [7]. Por su parte, el Análisis en Componentes Independientes (*Independent Component Analysis* ó ICA) da lugar a representaciones significativas cuando el conjunto de datos analizado responde a un modelo generativo de variables estadísticamente independientes [14]. Por otro lado, la Factorización en Matrices No Negativas (*Non-Negative Matrix Factorization* ó NMF) descompone linealmente conjuntos de datos bajo la restricción de la no negatividad de los mismos [9].

Pese a sus distintos enfoques, todas estas técnicas de representación tienen en común su capacidad para representar los datos observados sobre subespacios de dimensionalidad reducida, logrando así una compresión de los mismos, lo que resulta ventajoso desde el punto de vista de su almacenamiento y posterior procesamiento.

Cuando se considera la aplicación de este tipo de técnicas al análisis de secuencias AV, es preciso definir el espacio de datos sobre el que se opera. Así, la matriz **X** debe contener la información auditiva y visual de la secuencia analizada. Siguiendo el modelo propuesto en [14], en este trabajo se han representado los datos de audio

mediante el módulo de su espectro, mientras que la información visual consiste en el valor de los píxeles de cada fotograma de la secuencia de vídeo (ver figura 1).

En términos generales, se analiza una secuencia de N fotogramas de tamaño $M_v = M_v^x \times M_v^y$ píxeles. Para cada fotograma, se crea un vector audiovisual M-dimensional mediante la concatenación de *i)* el módulo de la transformada rápida de Fourier (*Fast Fourier Transform* ó FFT) de M_a puntos del segmento de audio asociado al fotograma, y *ii)* el vector M_v-dimensional resultante de reordenar cada fotograma en un vector (es decir, $M = M_a + M_v$). Así pues, la matriz $\mathbf{X}$ (de dimensiones $M \times N$) representa toda la secuencia AV.

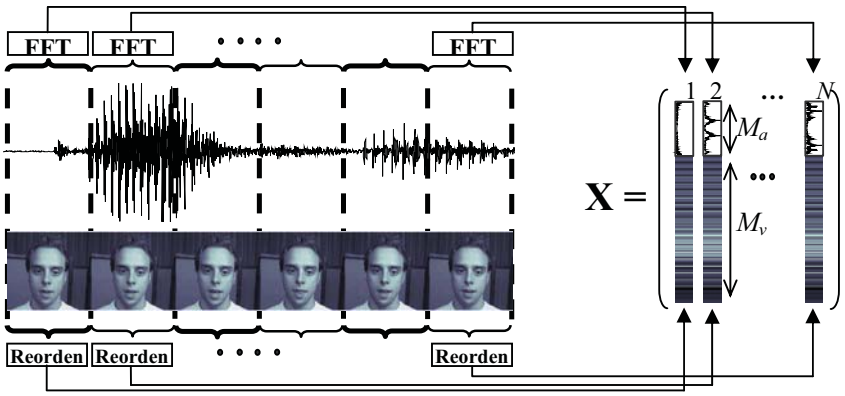

Fig. 1. Generación de la matriz $\mathbf{X}$ a partir de una secuencia audiovisual.

2.1 Extracción de bases AV mediante PCA

El Análisis en Componentes Principales [7] ofrece las direcciones principales de variación de los datos analizados, así como información relativa a la importancia de los mismos, siendo la descomposición en valores singulares (*Singular Value Decomposition* ó SVD) [8] una vía plausible para encontrarlas. La SVD es una técnica que factoriza cualquier matriz $\mathbf{X}$ de M filas, N columnas y rango K en el siguiente producto:

$$\mathbf{X}_{M\times N} = \mathbf{U}_{M\times K} \cdot \mathbf{\Sigma}_{K\times K} \cdot \left(\mathbf{V}_{N\times K}\right)^T + \mathbf{m}_{M\times 1} \cdot \mathbf{1}_{1\times N} \quad (1)$$

donde $\mathbf{U}$ y $\mathbf{V}$ son dos matrices ortonormales, $\mathbf{\Sigma}$ es una matriz diagonal, $\mathbf{m}$ es la columna media de $\mathbf{X}$ y $\mathbf{1}$ es un vector fila N-dimensional que sólo contiene unos. La matriz $\mathbf{U}$ contiene, en sus columnas, las componentes principales de $\mathbf{X}$, en la diagonal de $\mathbf{\Sigma}$, sus valores singulares y en las columnas $\mathbf{V}$, las componentes principales de $\mathbf{X}^T$.

La SVD se puede utilizar también para reducir la dimensionalidad de los datos [9] ya que proporciona la mejor aproximación $\mathbf{\Lambda}$ de rango R (siendo $R<K$) de una matriz $\mathbf{X}$ cualquiera, en el sentido de los mínimos cuadrados [4].

$$\mathbf{X}_{M\times N} \approx \mathbf{U}_{M\times R} \cdot \mathbf{\Sigma}_{R\times R} \cdot \left(\mathbf{V}_{N\times R}\right)^T + \mathbf{m}_{M\times 1} \cdot \mathbf{1}_{1\times N} = \mathbf{\Lambda} \quad \textbf{(2)}$$

Para representar un conjunto de datos audiovisual **X** en función de sus componentes principales, se realizará la SVD de rango R sobre éste (ecuación 2). Como resultado, se obtienen los R vectores base (columnas de **U**) y las proyecciones de cada columna de **X** sobre esta base en $\mathbf{C} = \mathbf{\Sigma V}^T$, que representan (por filas) los niveles de actividad de cada columna de **U**. Hay que destacar que las bases de **U** están formadas por dos tipos de información: sus primeras M_a dimensiones corresponden a la información auditiva, mientras que el resto (M_v) está asociado a información visual.

Para reducir el consumo de recursos de la aplicación de PCA mediante SVD, en [12] se propuso un esquema eficiente basado en el cálculo incremental de la SVD con actualización de la media. A partir de una inicialización que utiliza las primeras S columnas de **X** (ecuación 3), se obtienen las nuevas matrices $\mathbf{U}_{t+1}$, $\mathbf{\Sigma}_{t+1}$ y $\mathbf{V}_{t+1}$ y la nueva columna media $\mathbf{m}_{t+1}$ a partir de $\mathbf{U}_t$, $\mathbf{\Sigma}_t$, $\mathbf{V}_t$, $\mathbf{m}_t$ (ecuación 4) y S nuevas columnas de **X**, es decir, $\mathbf{x}_{tS+1}$, … $\mathbf{x}_{tS+S}$ (ecuación 5).

$$\mathbf{X}_0 = \left[\mathbf{x}_1 \ \ldots \mathbf{x}_S\right] = \mathbf{U}_0 \cdot \mathbf{\Sigma}_0 \cdot \left(\mathbf{V}_0\right)^T + \mathbf{m}_0 \cdot \mathbf{1} \quad \textbf{(3)}$$

$$\mathbf{X}_t = \mathbf{U}_t \cdot \mathbf{\Sigma}_t \cdot \left(\mathbf{V}_t\right)^T + \mathbf{m}_t \cdot \mathbf{1} \quad \textbf{(4)}$$

$$\mathbf{X}_{t+1} = \left[\mathbf{X}_t \quad \mathbf{x}_{tS+1} \ \ldots \mathbf{x}_{tS+S}\right] = \mathbf{U}_{t+1} \cdot \mathbf{\Sigma}_{t+1} \cdot \left(\mathbf{V}_{t+1}\right)^T + \mathbf{m}_{t+1} \cdot \mathbf{1} \quad \textbf{(5)}$$

Si el número de valores singulares se limita en cada paso de la iteración y se hace menor que el rango de **X**, las ecuaciones 3, 4 y 5 pasan a ser aproximaciones.

2.2 Extracción de bases AV mediante NMF

La Factorización en Matrices No Negativas se introdujo en [9] como técnica para aproximar una matriz no-negativa **X** (de dimensiones $M \times N$) mediante el producto de dos matrices no negativas: una matriz **W** ($M \times R$) y una matriz **H** ($R \times N$).

$$\mathbf{X} \approx \mathbf{W} \cdot \mathbf{H} = \mathbf{\Lambda} \quad \textbf{(6)}$$

Al igual que ICA, NMF asume la existencia de un modelo generativo. Sin embargo, NMF considera que el conjunto de datos multidimensional observado es el resultado de la suma ponderada de R fuentes (o vectores base) no negativas. De acuerdo a este modelo generativo, los R vectores base no negativos están contenidos en las columnas de **W** y los pesos asociados a cada vector base se encuentran en las filas de **H**. La matriz $\mathbf{\Lambda}$ denota la aproximación de la secuencia **X**.

En el contexto del análisis AV, el contenido de las matrices aproximadoras **W** y **H** debe ser interpretado cuidadosamente. Las primeras M_a dimensiones de la columna i-ésima de **W** describen el módulo del espectro de la parte auditiva de la i-ésima componente de la base AV. Sus M_v dimensiones restantes generan, después del reordenamiento adecuado, el modo visual de la i-ésima fuente AV. Por otra parte, la

fila *i*-ésima de **H** describe el nivel de actividad de la *i*-ésima fuente AV a lo largo de los *N* fotogramas de la secuencia audiovisual.

Desde el punto de vista algorítmico, la aproximación de la ecuación 6 puede realizarse mediante la optimización de una función de coste proporcional al error cuadrático medio (ECM) de reconstrucción (ecuación 7) [10]. Ésta se optimiza iterativamente mediante la aplicación de las reglas multiplicativas de actualización propuestas en [10], donde ⊗ denota el producto de Hadamard, las divisiones se calculan elemento a elemento y *T* denota transposición matricial (ecuación 8).

$$D_{\text{ECM}} = \left\| \mathbf{X} - \mathbf{\Lambda} \right\|^2 \tag{7}$$

$$\mathbf{H} = \mathbf{H} \otimes \frac{\mathbf{W}^T \cdot \mathbf{X}}{\mathbf{W}^T \cdot \mathbf{\Lambda}} \quad , \quad \mathbf{W} = \mathbf{W} \otimes \frac{\mathbf{X} \cdot \mathbf{H}^T}{\mathbf{\Lambda} \cdot \mathbf{H}^T} \tag{8}$$

3 Síntesis AV a partir de las bases extraídas

A causa de su alto contenido semántico (ver sección 4.1), las bases AV derivadas se prestan a ser usadas con propósitos sintéticos. Además, la naturaleza lineal de las descomposiciones PCA y NMF sugiere que la síntesis audiovisual puede ser llevada a cabo con un coste computacional razonable.

Sin embargo, sintetizar una secuencia a partir de un conjunto de bases AV requiere tener en cuenta una serie de consideraciones. En primer lugar, la síntesis del modo visual consiste en reordenar las M_v últimas dimensiones de los vectores base (columnas de **U** –bases PCA– ó **W** –bases NMF–) en matrices de dimensiones $M_v = M_v^x \times M_v^y$ (o fotogramas). Y en segundo lugar, sintetizar el modo auditivo implica el cómputo de una FFT inversa, la cual, además del módulo espectral (contenido en las M_a primeras dimensiones de las columnas de **U** ó **W**), requiere también de la respuesta de fase. En este trabajo, este asunto se resuelve mediante el uso de la fase del audio original [2].

Teniendo en cuenta dichas consideraciones, se pueden crear secuencias AV sintéticas mediante la combinación lineal de un conjunto determinado de vectores base. Es más, dado que las filas de la matriz $\mathbf{\Sigma V}^T$ ó **H** (según corresponda) describen la activación temporal de las componentes de las bases AV, éstas se pueden usar para segmentar la secuencia en relación a las bases AV extraídas. Así, una vez segmentada cada fuente AV, éstas pueden ser combinadas a voluntad para generar cualquier secuencia AV sintética. Por ejemplo, se puede realizar una extracción de fuentes AV si la síntesis se lleva a cabo mediante la selección de una única componente de las bases AV [14].

4 Experimentos

La viabilidad de los métodos propuestos ha sido evaluada y comparada mediante unos experimentos realizados sobre una secuencia de vídeo de un locutor pronunciando las cinco vocales castellanas **/aeiou/**. Las frecuencias de muestreo son 24 fotogramas/segundo para el vídeo y 11'025 kHz para el audio. Se ha empleado el algoritmo de cálculo incremental de la SVD propuesto en [12] y la implementación del algoritmo NMF basado en la métrica ECM extraído de NMFPACK [6].

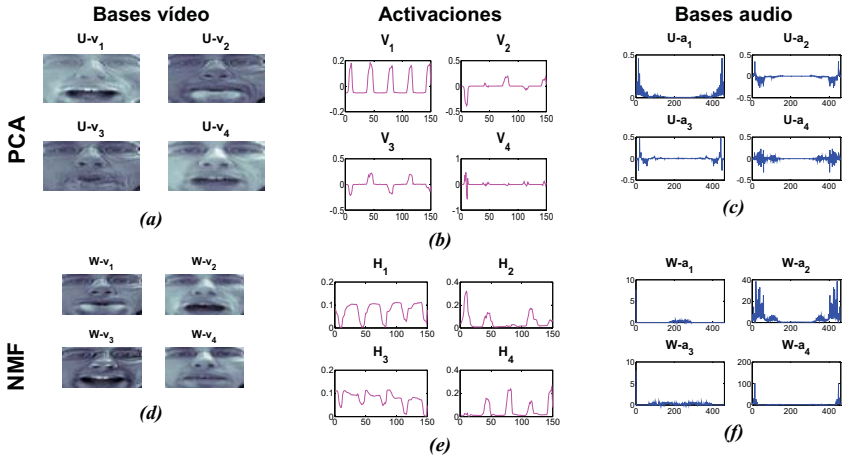

Fig. 2. *Bases audiovisuales y sus activaciones temporales correspondientes obtenidas mediante PCA y NMF para rango R* = 4.

4.1 Análisis audiovisual

En este experimento se visualizan las bases AV obtenidas mediante PCA y NMF con rango $R = 4$. Las bases AV y activaciones temporales resultantes se muestran en la figura 2. En lo referente a las bases PCA (fila superior), la figura 2a revela que las componentes visuales de las bases corresponden a apariencias faciales pronunciando distintas partes de la secuencia. La interpretación de dichas bases visuales se simplifica si se observan sus activaciones temporales (figura 2b). Por ejemplo, se aprecia que la primera componente de la base visual ($\mathbf{U}$-$\mathbf{v}_1$) se activa en cada pronunciación –no en vano muestra la boca abierta–, mientras que la tercera componente visual ($\mathbf{U}$-$\mathbf{v}_3$) es aditiva en la pronunciación de la /e/ y la /o/, sustractiva en el resto –ya que toma valores negativos en estos casos– e inactiva durante los silencios. Por su parte, las componentes auditivas de las bases (figura 2c), al tomar valores positivos y negativos, son difícilmente asociables a espectros de habla.

Por otro lado, la observación de las bases AV obtenidas por NMF (fila inferior de la figura 2), revela que la primera componente de la base visual ($\mathbf{W}$-$\mathbf{v}_1$) es activa durante los silencios –véase como el contenido espectral de la base auditiva asociada,

W-$\mathbf{a}_1$, es casi nulo–, mientras que el segundo vector base visual (**W**-$\mathbf{v}_2$) es activo durante la pronunciación de las vocales abiertas (/a/, /e/ y /o/).

En resumen, se aprecia que las bases NMF tienen un mayor contenido semántico que las obtenidas mediante PCA, o cuanto menos, su interpretabilidad es más sencilla. Esta característica, especialmente interesante para la segmentación de la secuencia AV analizada, se acentúa al incrementar la dimensionalidad del análisis *R*.

4.2 Evaluación objetiva de la resíntesis audiovisual

Como primer paso hacia la aplicación de las bases AV obtenidas mediante PCA y NMF con propósitos de síntesis audiovisual, la secuencia analizada fue resintetizada en su totalidad a partir de las bases AV obtenidas mediante ambas técnicas, generándose la secuencia aproximada $\mathbf{\Lambda}$ (ver ecuaciones 2 y 6).

En este experimento se comparan objetivamente la secuencia resintetizada $\mathbf{\Lambda}$ y la secuencia original **X** mediante el cálculo del error cuadrático medio normalizado (ECMN) entre los modos audio y vídeo de ambas secuencias. Con el fin de analizar el impacto de la dimensionalidad del análisis (*R*) en la calidad de la síntesis, ésta fue variada desde 2 hasta 40 en el proceso de resíntesis.

La figura 3a muestra el ECMN del modo visual. Se aprecia que, en este modo, NMF presenta un error menor que PCA, aunque este diferencial se reduce al aumentar *R*. Este efecto se debe al carácter incremental del cálculo de la SVD utilizado, el cual obtiene, para valores pequeños de *R*, una reconstrucción deficiente de las características visuales novedosas que aparecen en la secuencia a lo largo del tiempo.

En cuanto al audio (figura 3b), PCA obtiene una mejor aproximación que NMF en todo el rango de valores de *R* estudiado. Ésto resulta sorprendente, máxime cuando las componentes auditivas de las bases PCA toman tanto valores positivos como negativos y la representación original del audio es el módulo de la FFT.

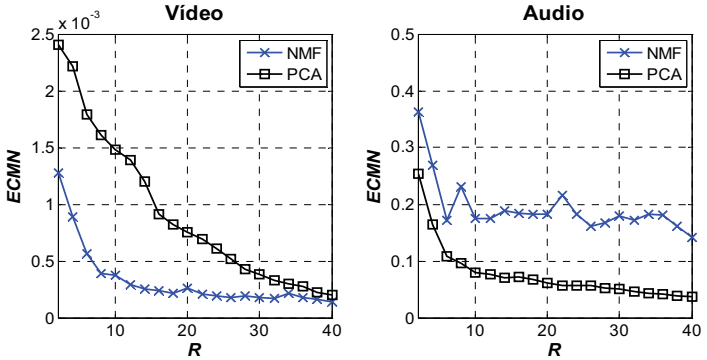

Fig. 3. Error cuadrático medio normalizado entre *(a)* los vídeos original y resintetizado, y *(b)* los audios original y resintetizado en función de la dimensionalidad del análisis.

4.3 Evaluación subjetiva de la resíntesis audiovisual

Este experimento tiene como objetivo evaluar la calidad de las secuencias resintetizadas desde una óptica subjetiva. A tal efecto, se sometió a un conjunto de 13 usuarios a dos pruebas: un test de opinión y un test de preferencia. Por ello, se prepararon diferentes secuencias que combinaban los modos visual y auditivo originales y resintetizados obtenidos por PCA y NMF con rangos $R = \{4, 10, 40\}$.

En el test de opinión se presentaron sucesivos vídeos a cada usuario y éste debía calificarlos con un *Mean Opinion Score* (MOS) entre 5 (calidad máxima) y 1 (calidad mínima). Con el fin de uniformizar los criterios de los distintos usuarios, inicialmente se mostraban ejemplos de secuencias calificadas con los dos valores extremos de la escala MOS. Los resultados de esta prueba (figura 4) muestran que los usuarios perciben una mayor calidad en la síntesis visual por NMF, mientras que valoran más positivamente la síntesis auditiva mediante PCA. Obsérvese que esto concuerda con los valores objetivos de ECMN mostrados en la figura 3. Por otra parte, la evaluación de la síntesis conjunta de ambos modos arroja un MOS mayor para PCA que para NMF. Este resultado demuestra la importancia de la calidad del audio resintetizado en la evaluación global de la secuencia AV sintética. Por último, a lo largo de toda la evaluación se comprueba el aumento de MOS al incrementarse la dimensionalidad del análisis R, lo que concuerda con la evolución decreciente del error ECMN mostrado en la figura 3.

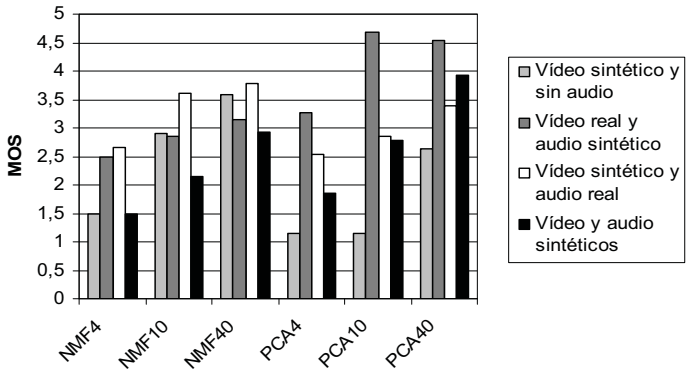

Fig. 4. Resultados del test *Mean Opinion Score* (MOS) sobre distintas secuencias sintetizadas mediante PCA y NMF con rangos R iguales a 4, 10 y 40.

Por otro lado, se realizó un test de preferencia mostrando parejas de vídeos, ante los que el usuario debía escoger el que considerase más real, ofreciéndosele también una tercera opción de indistinción. El objetivo de esta prueba es realizar una comparación directa entre ambas técnicas de síntesis AV (PCA y NMF), presentando secuencias que tienen el vídeo y/o el audio real o sintetizado. Los resultados (figura 5) indican, de nuevo, una clara preferencia por el método NMF al sintetizar vídeo y por PCA al sintetizar audio. Sin embargo, este nivel de preferencia se reduce notablemente al comparar directamente las síntesis audiovisuales conjuntas obtenidas

mediante ambas técnicas, aunque aún se manifiesta una ligera tendencia de preferencia hacia PCA.

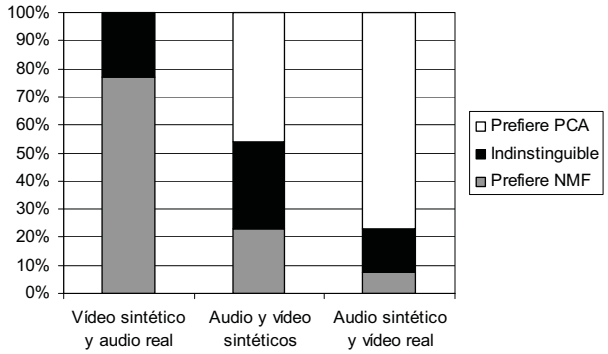

Fig. 5. Resultados del test de preferencia en síntesis mediante PCA y NMF de rango $R = 40$.

5 Conclusiones

En este trabajo se han aplicado y comparado los métodos de análisis en componentes principales y descomposiciones lineales con restricciones de no negatividad en el contexto del análisis y síntesis AV de secuencias de vídeo facial con voz. Mientras que la mayoría de aproximaciones previas al análisis de secuencias AV se han centrado en la detección de dependencias audiovisuales, nuestra propuesta se orienta hacia el uso de las bases extraídas con propósitos de síntesis AV en el marco del desarrollo de interfaces que faciliten una IPO lo más natural posible.

El experimento de análisis audiovisual realizado muestra diferencias significativas en las bases obtenidas por los métodos PCA y NMF. En concreto, las bases NMF poseen una interpretación física más intuitiva que las ofrecidas por PCA, probablemente debido al carácter aditivo de la descomposición NMF. Por este mismo motivo, las activaciones temporales de las bases NMF parecen ser más fácilmente segmentables, lo que puede ser de gran utilidad de cara a futuras investigaciones centradas en síntesis AV.

La evaluación de las secuencias resintetizadas ha arrojado resultados muy interesantes. Desde un punto de vista objetivo, el error cuadrático medio normalizado cometido al aproximar los modos de audio y vídeo varía en función del método utilizado: mientras NMF presenta un menor error en la representación de la información visual, PCA obtiene mejores resultados en el modo auditivo, pese a que la no negatividad inherente a NMF pudiese inducir a pensar lo contrario.

Los resultados obtenidos en las pruebas de evaluación subjetiva corroboran el análisis objetivo del error previamente realizado: es decir, los usuarios han considerado que la síntesis del modo visual obtenida mediante NMF supera a PCA. Por el contrario, la calidad del audio sintetizado a través de PCA resulta más alta que la obtenida mediante NMF. No obstante, al someter a los usuarios a un test de

preferencia sobre las secuencias AV resintetizadas, cabe destacar el notable nivel de indecisión observado al comparar directamente ambos métodos. Como línea de trabajo futuro, se considera que el empleo de representaciones de audio que incluyan la información de fase de la señal puede constituir un primer paso hacia la mejora, por otro lado necesaria, de la calidad del audio sintético.

Referencias

1. Butz, T. y Thiran, J.P. From error probability to information theoretic (multi-modal) signal processing. Signal Processing, 85(5):875–902 (2005)
2. Casey, M.A. y Westner, A. Separation of Mixed Audio Sources By Independent Subspace Analysis. Proceedings of the International Computer Music Conference (ICMC 2000), 154–161. Berlín, Alemania (2000)
3. Fisher III, J.W., Darrell, T., Freeman, T.W. y Viola, P. Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. Advances in Neural Information Processing Systems, vol. 14 (2000)
4. Golub, G. y Loan, C.V. Matrix Computations. The Johns Hopkins University Press, 1996
5. Hershey, J. y Movellan, J. Audio-Vision: Using Audio-Visual Synchrony to Locate Sounds. Advances in Neural Information Processing Systems, vol. 12 (1999)
6. Hoyer, P.O. Non-Negative Matrix Factorization with Sparseness Constraints. Journal of Machine Learning Research, 5:1457–1469 (2004)
7. Jolliffe, I. Principal Component Analysis. New York Springler-Verlag, 1986.
8. Kirby, M.. Geometric Data analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns. John Wiley & Sons Inc., Nueva York, 2001.
9. Lee, D.D. y Seung, H.S. Learning the Parts of Objects by Non-Negative Matrix Factorization. Nature, 401, 788–791 (1999)
10. Lee, D.D. y Seung, H.S. Algorithms for Non-Negative Matrix Factorization. Advances in Neural Information Processing Systems, vol. 13 (2000)
11. Melenchón, J., Iriondo, I., Socoró, J.C. y Martínez, E. Lip Animation of a Personalized Facial Model from Auditory Speech. Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2003), 187–190. Darmstadt, Alemania (2003)
12. Melenchón, J., Meler, L. y Iriondo, I.. On-the-fly training. Articulated Models and Deformable Objects, LNCS vol. 3179, pp. 146–153, Palma de Mallorca, España (2004)
13. Slaney, M. y Covell, M. Facesync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. Advances in Neural Information Processing Systems, vol. 13 (2000)
14. Smaragdis, P. y Casey, M. Audio/Visual Independent Components. Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Source Separation (ICA 2003), 709–714. Nara, Japón (2003)