El control objetivo de la implicación de los informantes en el estudio del habla emocionada.

Publicado en: Procesamiento del Lenguaje Natural, revista nº 23, Alicante, septiembre de 1998, ISSN: 1135-5948, (pp.24-31).

Ángel Rodríguez Bravo
Patricia Lázaro
Norminanda Montoya
Josep Mª Blanco
Dolors Bernadas
Josep Manel Oliver

Ludovico Longhi
Mireia Gonzalez
Universidad Autónoma de Barcelona
Dto. De Comunicación Audiovisual y Publicidad
Edificio I, 08193 Bellaterra (Barcelona)
Arbravo@intercom.es

Resumen

En las investigaciones sobre la influencia de las emociones en el habla aparece sistemáticamente el problema de cómo objetivar el estado emocional de los informantes (locutores). Así, para construir los córpora (o bases de datos) se suele solicitar a uno o varios actores una serie de interpretaciones, proporcionándoles los textos y una serie de indicaciones más o menos completas sobre el tipo de estado emocional que han de interpretar. Luego, se asume ya, normalmente sin más control, que la interpretación del actor o actores contienen la "verdad" sonora.

Si para cualquier buen aficionado al teatro o al cine no es ningún secreto que la interpretación de los actores con frecuencia dista mucho de ser creíble, no parece lógico que se asuma como válida para ser analizada acústicamente la interpretación de cualquier actor sin ningún otro tipo de control objetivo.

En esta comunicación se exponen dos métodos para objetivar el estado emocional de los actores informantes: *a) el análisis de dos parámetros fisiológicos* (concretamente ritmo cardiaco y tensión arterial), y *b) el desarrollo de un test perceptivo a sujetos experimentales.* Finalmente, se defiende como conclusión central la necesidad de diferenciar entre:

. Reconocimiento de la emoción,
. Verosimilitud de la emoción,
. Vinculación entre parámetros fisiológicos y estado emocional del hablante.

Los métodos que se presentan han sido diseñados y aplicados en el marco del proyecto *"Modelización acústica de la expresión emocional en el español"* financiado por la DGICYT (PB94-0732).

1. Introducción.

La variabilidad del habla en su sentido más amplio es, sin lugar a dudas, el problema fundamental que se opone todavía a un avance rápido en el procesamiento automático del lenguaje natural. La extraordinaria riqueza informativa que hay codificada en la polifonía de cualquier discurso oral humano está aun muy escasamente formalizada y, en consecuencia, una de las cuestiones centrales de este campo sigue siendo la modelización y la selección de parámetros adecuados y eficientes para representar la información del discurso oral.

A nuestro modo de ver, la solución de este problema ha de pasar por una línea de investigación básica que sea capar de modelizar la variabilidad del habla desde una perspectiva que sea alternativa y a la vez complementaria a la que aportan los sistemas de extracción automatizada de parámetros basada en los modelos de Markov, las redes neuronales o las medidas de correlación. De hecho, la mayoría de trabajos publicados en los últimos años, entre los que podemos citar como significativos: ROSENBERG y cols.: 1990; MARTÍNEZ, F.: 1991; GARCÍA-SÁNCHEZ, A. y cols.: 1993; MATSUI, T. and FURUI, S.: 1992; SAVIC, M and SORENSEN, J.: 1992; MARTIN, F. y cols.: 1993; GALES M.J.F. and YOUNG, S.J.: 1993; KATAGIRI, S. and JUANG, B.H.: 1993; FARRELL, R.J.: 1994;

ROSE y cols.: 1994; MATSUI, T. and FURUI, S.: 1994; ARTIERES, T. and GALLINARI, P.: 1995; LIOU, H.S. and MAMMONE, R.J.: 1995; MATSUI y cols.: 1995; SETLUR, A and JACOBS, T.: 1995, se centran en este tipo de planteamientos que son ya clásicos, pero que siguen sin dar una solución definitiva a la localización de parámetros eficientes.

Es cierto que actualmente existen ya en el mercado sistemas de reconocimiento de voz que consiguen un nivel de eficacia relativamente satisfactorio, es, por ejemplo, el caso de "*Simply Speaking Golg"*, *"Via voice"* y "*Via Voice Gold*" distribuidos por IBM; o de "*Dragon Dictate NaturallySpeaking"* que distribuye Naga; o los traductores texto-voz como el desarrollado por Telefónica. No obstante, todos estos sistemas siguen limitados a lo que podríamos denominar como *"habla neutra"*. Es decir, solo reconocen, o sintetizan un tipo de habla inexpresiva, que no admite ninguna de las variaciones habituales y características del habla humana que son independientes y paralelas a la estructura de la lengua. En el momento en que se intenta introducir repentinamente en cualquiera de ellos aspectos como cambios de locutor, matizaciones sonoras enfáticas, cambios en el estado emocional, etc. todos, sin ninguna excepción, dejan de ser eficientes.

En los últimos años se ha abierto una nueva corriente que toma como objeto de estudio la variabilidad expresiva del habla. Es decir, que estudia las sucesivas desviaciones de los parámetros acústicos centrales que sufre una misma estructura lingüística, cada vez que es repetida con una actitud psicológica distinta, desde diferentes estados emocionales y afectivos, por locutores diversos, etc. (RODRIGUEZ: 1989; GARRIDO: 1990; ENGSTRAD: 1992; AGUILAR, BLECUA, MACHUCA y MARIN: 1993; GARRIDO, LLISTERRI, MOTA, Y RIOS: 1993; VROOMEN, COLLIER y MOZZICONACCI: 1993; FONAGY: 1983; GIMÉNEZ:1986; SCHERER y EKMAN: 1984; SCHERER: 1988, 1995; BANSE y SHERER: 1996; LEINONEN y col. 1997; PROTOPAPAS y LIEBERMAN: 1997; PERROT,J.: 1997).

La investigación que presentamos aquí se inscribe en esa corriente que intenta explorar y formalizar las distintas dimensiones acústicas que comporta la variabilidad expresiva del lenguaje natural. Lógicamente, las posibilidades de agotar en una investigación la polifonía del habla son remotas, en consecuencia, nuestro estudio acota y se centra en uno de los aspectos de la expresión oral que es, sin duda, de los más relevantes: el habla emocionada.

De hecho, hace ya unas dos décadas que se inició el trabajo científico sistemático sobre la influencia de las emociones en el lenguaje oral en su sentido más amplio, podemos citar, por ejemplo a KNAPP: 1980; SHMIDT-ATZERT: 1985; RENCHLIN: 1987; MAYOR:1988; SCHERER: 1988; DEMANY y SORIN: 1989. En el caso de España probablemente uno de los trabajos más interesantes en esta línea fue el desarrollado por A. Giménez Fernández (GIMENEZ: 1986). Este tipo de investigaciones sobre el habla emocionada tiene un momento especialmente álgido a finales de los años 80 y principios de los 90. Es de destacar, por ejemplo, la excelente revisión que publicaron I.R. Murray y J.L. Arnott sobre las investigaciones experimentales desarrolladas en esa etapa en torno a la expresión acústica de las emociones y sus posibilidades de implementación en los sistemas de síntesis (Cfr. MURRAY y ARNOTT: 1993).

No obstante, el interés por estudiar el habla emocionada parece haber disminuido en los últimos años, probablemente debido a la falta de eficiencia en la obtención de los resultados que comporta la construcción de bases de datos con habla emocionada sin las suficientes garantías metodológicas.

2. Los problemas cásicos en los corpora de habla emocionada.

Para construir los córpora (o bases de datos) en las investigaciones dedicadas a buscar parámetros acústicos que puedan reflejar la influencia de las emociones en el habla, se suele recurrir a actores que serán los informantes que expresen los distintos estados emocionales. Una vez seleccionado el actor (o actores) los investigadores le dan una serie de indicaciones más o menos completas sobre las emociones que ha de interpretar, luego, se procede directamente a grabar el corpus. A partir de este momento se asume ya, sin ningún otro tipo de control metodológico objetivo, que la interpretación del actor (o los actores) contienen la "verdad" sonora.

Si para cualquier buen aficionado al teatro o al cine no es ningún secreto que las interpretaciones de los actores con frecuencia distan mucho de ser creíbles, no parece lógico

que sea tomada como patrón acústico de referencia una interpretación supuestamente emocionada de la voz, por el simple hecho de que provenga de un actor profesional. Obviamente, cuando una interpretación oral es percibida como no creíble, ello implica que en esa voz sobra o falta algo respecto al patrón acústico registrado en nuestra memoria auditiva. Y es importante tener en cuenta que ese patrón acústico no solo proviene de nuestra experiencia auditiva en la comunicación con los demás, sino de la experiencia de escuchar la propia voz cuando vivimos y sentimos cada una de nuestras emociones (esta última es, sin duda, la situación de referencia más sincera y realista posible).

Es cierto que el investigador observará siempre el resultado de las interpretaciones del actor o los actores informantes y acordará con ellos cuales han sido las mejores "actuaciones" para conservarlas en el corpus. Pero también lo es que el estudioso, normalmente, está en inferioridad de condiciones frente a un actor profesional para valorar la calidad de la interpretación, y eso hace que sea la opinión del propio actor la que predomine en los criterios de selección. No obstante, el problema principal que subyace en este método para la construcción del corpus de habla emocionada es, en realidad, la saturación semántica que inevitablemente va a sobrevenir a los investigadores durante el proceso de preparación y desarrollo de la grabación. Nos estamos refiriendo a la pérdida de capacidad por parte del equipo de investigadores para valorar con objetividad si una interpretación emocionada es correcta o no, después de haber preparado los textos, de haberlos estudiado e interpretado en numerosas ocasiones y, luego, haber negociado, orientado, sugerido, corregido, grabado…, con el actor cada una de las interpretaciones infinidad de veces.

Finalmente, mientras se prepara, en este tipo de corpus surge sistemáticamente el problema de tener que decidir si es valido, o no, que el actor informante finja sus emociones durante la interpretación. Cuando se solicita al actor que trabaje intentando experimentar realmente las emociones propuestas, éste suele responder que él no siempre actúa desde sus propias emociones, sino que con frecuencia utiliza distintas técnicas para expresarlas sin sentirlas realmente.

En suma, las condiciones suelen ser las siguientes: 1) *predominio del criterio profesional del actor*; 2) *saturación semántica de los investigadores*; 3) *posibilidad de que el actor informante finja sus interpretaciones emocionadas*. En esta situación lo que conseguimos, en realidad, es un corpus muy incompleto basado simplemente en que un actor consiga que sus actitudes emocionadas sean *reconocibles* por un grupo de investigadores, normalmente no entrenados en las técnicas de interpretación, que además, están saturados de escuchar infinidad de veces las mismas frases con sutiles cambios sonoros. A nuestro modo de ver, en estas condiciones difícilmente se puede garantizar que un corpus contenga realmente suficientes muestras de habla emocionada completa y bien construida en todas sus dimensiones acústicas.

Obviamente, si el corpus de referencia a partir del cual esperamos modelizar acústicamente las emociones no contiene aquello que buscamos, o solo lo contiene en parte, es difícil que podamos obtener de él modelos realmente representativos a partir de su análisis acústico. En consecuencia, consideramos que una de las razones fundamentales que está dificultando la obtención de parámetros acústicos eficientes para representar las emociones del habla es la inadecuada construcción de los córpora.

3. ¿Es posible garantizar la objetividad en la construcción de un corpus oral de habla emocionada?

Ciertamente, creemos que las investigaciones sobre psicología de las emociones desarrolladas en la última década aportan conocimientos concretos que, efectivamente, permiten estudiar con objetividad el grado de implicación real de un actor al interpretar locuciones emocionadas. Así que a continuación expondremos cuales son los puntos de partida teóricos en los que apoyamos nuestra afirmación.

Existe un acuerdo ampliamente generalizado entre los psicólogos en concebir la emoción como una reacción orgánica compuesta por tres componentes: 1) *experiencia subjetiva*, 2) *conducta expresiva*, y 3) *respuesta fisiológica* (Cfr. REEV, 1994:320). Así, las emociones no son exclusivamente experiencias subjetivas, sino que implican, también, cambios anatómicos, neurofisiológicos y endocrinos involuntarios, que se desencadenan al experimentar la emoción (Cfr. IZARD, 1982). Ese carácter fisiológico y no voluntario da a la

emoción un rasgo diferencial muy claro, por ejemplo, respecto al concepto más vago y difuso de *sentimiento*, en tanto que con los sentimientos los individuos no experimentan modificaciones somáticas (Cfr. DANCER, 1989:29); convirtiendo así las emociones en un fenómeno relativamente fácil de identificar en una investigación.

Estamos, pues, frente a un fenómeno cuyo carácter expresivo no es arbitrario, sino que está determinado por la propia fisiología del individuo que lo experimenta. Eso significa, lógicamente, que la observación de los cambios fisiológicos asociados a cada emoción pueden ser utilizados como indicadores para comprobar si durante cada locución un actor informante está experimentando realmente alguna emoción, o no. Sabemos, además, que cada emoción provoca una respuesta fisiológica concreta y que generalmente hay coherencia entre la experiencia emocional y su expresión (Cfr. EKMAN, 1983). De hecho, el componente expresivo de las emociones ha sido ya minuciosamente estudiado y formalizado en lo que se refiere a la conducta facial, y se ha aceptado de forma generalizada que la expresión de las emociones básicas es reconocida universalmente a través de mímicas faciales específicas. Es cierto que existe la posibilidad de modificar voluntariamente la expresión al margen de la emoción que pueda estarse experimentando. No obstante, si lo que estamos buscando es una vinculación objetiva entre la expresión del sujeto emocionado y la emoción misma, esto es fácilmente resoluble solicitando a los informantes que no repriman ni intenten manipular su actitud y evitando, a la vez, aceptar como informantes a sujetos inexpresivos.

Disponemos, pues, de investigaciones que nos indican qué tipo de respuestas fisiológicas sufre un sujeto al experimentar una emoción (Cfr. REEV, 1994:324 y COSNIER, 1994:29) comprobándose, por ejemplo, diferencias en el ritmo cardiaco, la presión arterial y la temperatura de la piel según la emoción experimentada: "*ver un tigre aumenta el ritmo cardiaco y la presión sanguínea, perder dinero reduce el ritmo cardiaco y la presión sanguínea" (…) "el ritmo cardiaco aumentaba en la rabia, el miedo y la angustia pero cambiaba muy poco para la alegría la sorpresa y el asco*" (REEVE, 1994:323-324). Y disponemos, también, de un acuerdo muy generalizado entre los distintos estudiosos de la emoción respecto a que existe sólo un número muy reducido de emociones básicas (Cfr. IZARD, 1977; DANTZER, 1988:29; EKMAN, 1983; REEVE, 1994:373). En lo que ya no existe acuerdo concreto es en el número de éstas (que oscila entre 5 y 10) ni en los términos utilizados para denominarlas.

Resumiendo, los conocimientos que aporta actualmente la psicología de las emociones nos permiten establecer una serie de estados emocionales fundamentales o básicos y nos indican que cada una de ellos esta asociado a un tipo de conciencia interna, de conducta expresiva y de modificaciones fisiológicas. En consecuencia, si conseguimos observar en los actores informantes alguno de los indicadores fisiológicos que se alteran con la aparición de las distintas emociones, también podremos saber objetivamente si el actor ha experimentado realmente emoción mientras interpreta cada uno de los discursos, o si solo la está fingiendo. Así pues, de todo esto se desprende que es posible objetivar los estados emocionales de un informante si se dispone de los instrumentos de control adecuados.

4. La obtención de datos, de variables fisiológicas vinculadas a la emoción.

Nuestro corpus tenía que estar construido a partir de las locuciones de 8 actores, 4 hombres y 4 mujeres. Cada uno de ellos tenía que interpretar dos textos distintos en diferentes estados emocionales y con tres distintos grados de intensidad en cada estado emocional. Concretamente, en cada una de las lecturas emocionadas tenía que quedar reflejada una de las siguientes emociones: *sorpresa*, *alegría*, *deseo*, *rabia*, *tristeza*, *asco* y *miedo*. El objetivo final era que el corpus contuviera cada uno de esos estados emocionales expresados con suficiente naturalidad y verosimilitud, para que luego fuese posible analizarlos y modelizarlos acústicamente.

A continuación explicaremos con detalle como se desarrolló el proceso de producción del corpus.

En primer lugar, el actor debía interpretar un texto en actitud neutra (no emocionada). El objetivo de esta primera lectura era familiarizar al actor con el texto y disponer de sus datos fisiológicos en un estado supuestamente no emocionado (mas tarde descubriríamos que en esta primera lectura el actor solía estar en gran tensión). A continuación se solicitaba ya al

actor que interpretase la primera emoción, por ejemplo la *sorpresa*, en primer lugar con poca intensidad emocional, luego con una intensidad emocional media y, finalmente, con una intensidad emocional alta. Para facilitar al actor su trabajo de puesta en situación, una vez aclarada la emoción que tenía que interpretar se le proponía una frase concreta para sugerirle cada grado de intensidad emotiva. Así, en el caso de la *sorpresa* las frases de sugerencia fueron las siguientes: para proponer poca intensidad se le dijo al actor "*algo te extraña*", para la intensidad media se le dijo "*has quedado pasmado*", y para sugerirle una intensidad emocional muy fuerte se le propuso "*estás completamente desconcertado*". Lógicamente, una vez planteada cada situación dramática se iniciaba una grabación sonora, con lo que todas las interpretaciones quedaron registradas en cinta magnetofónica. Este proceso se repetía con todas las emociones, proponiendo siempre al actor frases de sugerencia distintas y adecuadas a cada emoción concreta. Por ejemplo, en el caso de la *alegría* las frases de sugerencia fueron: "*te sientes contento*", "*estas entusiasmado*" y "*te has puesto completamente eufórico*". Posteriormente, se repetía de nuevo todo este ciclo interpretativo con un segundo texto, de modo que el corpus quedaría compuesto finalmente por las interpretaciones de 2 textos, por parte de 8 actores, interpretando cada texto con 7 emociones básicas y en tres grados de intensidad distintos cada una de ellas. Es decir, dispondríamos de un corpus con 2 x 8 x 7 x 3 = 336 interpretaciones emocionadas distintas.

Probablemente, los actores habrán conseguido emocionarse realmente sólo en algunas de estas 336 interpretaciones, el resto serán interpretaciones puramente "técnicas", es decir, fingidas. Debíamos, pues, diseñar un método basado en la observación de varios indicadores fisiológicos de los actores, que nos permitiese decidir de forma objetiva qué interpretaciones estaban realmente vinculadas a alguna emoción y qué otras no lo estaban. Dicho de otro modo, necesitábamos un método de control objetivo de la implicación emocional de los actores; de esa manera, una vez desarrollado el análisis acústico, sería posible vincular con garantías metodológicas suficientes ciertos parámetros acústicos de la voz con estados emocionales concretos.

Para conseguir ese control, basándonos en los conocimientos sobre psicología de las emociones expuestos más arriba, decidimos observar durante cada interpretación del actor algunas constantes fisiológicas que fuesen capaces de revelarnos objetivamente su estado emocional, concretamente se decidió estudiar el ritmo cardiaco y la presión arterial. No obstante, quedarían aun dos cuestiones por resolver: ¿Que ocurriría si los actores, a pesar de emocionarse realmente, por alguna razón no llegaban a expresar su emoción en la voz?. O al revés: ¿Como podríamos saber si un actor es capaz de engañarnos y expresar en su voz con total eficacia una emoción sin llegar en realidad a experimentarla fisiológicamente?

El único modo de resolver estas dos cuestiones era procediendo a una validación posterior del corpus, es decir, haciéndoselo escuchar y juzgar a un grupo amplio de sujetos experimentales no implicados en el proyecto. Estos sujetos clasificarían finalmente las interpretaciones y nos dirían qué voces expresaban con precisión y verosimilitud una emoción y qué otras no. Finalmente, la validación nos permitiría decidir, también, si las emociones, tal como las entiende la psicología, son o no son, objetivamente observables en el sonido de la voz. Esta segunda etapa de la validación del corpus se describe con más detalle un poco más abajo.

Los datos sobre la presión arterial y ritmo cardiaco de cada actor se obtuvieron utilizando tensiómetros *"Omron R1"* modelo HEM-601 R1 distribuido por la empresa Omrom Corporation. La toma de la presión y pulso se realizaba inmediatamente después de que el actor finalizara cada una de sus interpretaciones, teniendo la precaución de espaciar las lecturas de modo que entre una toma y otra transcurriese un mínimo de 2,5 minutos, con objeto de garantizar la recuperación del volumen normal de las arterias. Una vez medidas las constantes fisiológicas, las cifras obtenidas eran introducidas en una ficha sobre el actor, asociándolas mediante un código numérico a cada estado emocional propuesto y a la interpretación concreta que las había generado.

Se realizaron, en total, 42 tomas de la presión arterial y el ritmo cardiaco para cada uno de los 8 actores. Cada una de las mediciones aportaba tres datos: presión arterial sistólica en mmHg, presión arterial diastólica en mmHg y ritmo cardiaco en número de latidos por minuto).

4.1. Exploración y resultados de los datos fisiológicos.

Este tipo de información se muestra muy estrechamente asociada a cada individuo. Es decir, observamos que cada actor tenía sus propios niveles específicos de tensión arterial y ritmo cardiaco, y que los datos fisiológicos de cada uno de ellos resultaban muy homogéneos respecto al propio sujeto y, a la vez, claramente distintos a los de todos los demás.

Para resolver este problema optamos por comparar a cada sujeto consigo mismo. La forma de hacerlo fue calculando las constantes fisiológicas medias de cada actor. Así, a partir de la totalidad de los datos de cada informante, se obtuvieron tres cifras de referencia (presión sistólica media, presión diastólica media, y pulso medio) que fueron tomadas como punto de referencia para estudiar de qué modo quedaba afectado un actor al experimentar una emoción u otra. Este método permitía, luego, estudiar si el tipo de variación de las constantes fisiológicas que un informante experimentaba, por ejemplo al interpretar la alegría, era (o no era) común a la que experimentaban todos los demás.

Pudimos observar que, efectivamente, se manifestaban unas tendencias claras en varias emociones. Así, por ejemplo, pudimos constatar lo siguiente:

ALEGRIA: indujo a subir la p.sistólica en 5 de los 8 actores.
ASCO: indujo a bajar la p.sistólica en 5 de los 8 actores.
MIEDO: indujo a bajar la p.sistólica en 7 de los 8 actores.
SORPRESA: indujo a subir la p. sistólica y la p. diastólica en 6 de los 8 actores.
TRISTEZA: indujo a bajar el ritmo cardiaco en 5 de los 8 actores.

(No se observó ninguna tendencia clara para el DESEO ni para la RABIA)

Observamos, también, que frente a una emoción se producían dos respuestas posibles: 1) el locutor se excitaba y tendía a la agitación, lo cual provocaba una tendencia a subir su ritmo cardiaco y bajar su presión arterial; 2) el locutor se inhibía y tendía a la depresión, entonces tendía a desencadenarse la tendencia inversa, es decir, a bajarle el ritmo cardiaco y subirle la presión arterial.

Finalmente, se observó que aparecían variaciones especialmente acusadas de las variables fisiológicas en algunas interpretaciones concretas. Puesto que, teóricamente, las voces asociadas a estas características debían ser las "mejor" emocionadas, se procedió a estudiar cual había sido la valoración que habían hecho de ellas algo más de un millar de sujetos experimentales, desarrollando así la validación perceptiva del corpus que citábamos más arriba.

5. Validación perceptiva del corpus con oyentes no implicados.

La objetivación definitiva de nuestro corpus se realizó, pues, sometiendo cada una de las 336 interpretaciones emocionadas al juicio de más de 30 sujetos experimentales. Organizados en distintos grupos, se hizo escuchar a los sujetos experimentales una serie de interpretaciones. Tras cada interpretación, se solicitaba mediante un test a cada oyente que respondiese a las tres cuestiones siguientes: 1) debía concretar qué emoción o emociones reconocía en cada voz; 2) tenía que asignar un grado de verosimilitud al locutor; y, por último, 3) debía especificar si había llegado a emocionarse, o no, escuchando la interpretación y, si la respuesta era afirmativa, concretar en qué grado.

Este tipo de test nos permitió decidir con objetividad, es decir, con total independencia de los investigadores, qué interpretaciones contenían realmente informaciones acústicas concretas asociadas a las emociones. Y permitió, además, vincular las emociones que fueron reconocidas en cada voz, con las distintas tendencias observadas en las variaciones fisiológicas. En esta última fase de validación, comprobamos, efectivamente, que los oyentes consideraban bastante más verosímiles las interpretaciones que tenían asociados unos datos fisiológicos muy claramente desviados de sus promedios; y pudo observarse, también, que este fenómeno se producía con independencia de la emoción. No obstante, esta vinculación sistemática entre la desviación extrema de los datos fisiológicos con una mayor verosimilitud solo se cumplía en algunos locutores, en cambio en otros no fue así. Probablemente esto se debe la distinta capacidad de los

actores que actuaron como informantes de nuestro corpus, para autoinducir su estado psicológico durante las interpretaciones en un estado emocional objetivo y real.

6. Conclusiones.

En tanto que la investigación que presentamos está aun sin concluir al no haberse terminado aun la etapa de análisis acústico, no podemos mostrar todavía si los parámetros conseguidos después de aplicar este método de control y validación del corpus mejoran en precisión y eficiencia. No obstante, los resultados obtenidos hasta ahora reflejan claramente la utilidad de los instrumentos desarrollados para discriminar entre "buenas" y "malas" interpretaciones de la emoción por parte de los informantes.

Quizás una de las conclusiones más interesantes de este estudio es que el propio hecho de utilizar actores como informantes dificulta el control objetivo de sus estados emocionales y, en consecuencia, distorsiona los modelos que puedan ser obtenidos a partir del análisis acústico de sus interpretaciones. En tanto que un actor es un profesional preparado para fingir la expresión de emociones, eso desencadena un nivel alto de confusión entre los resultados expresivos de las voces con la emoción fingida y las voces emitidas durante un estado emocional real. Así, interpretaciones que son reconocibles como emocionadas (consiguen reproducir algunos rasgos acústicos de la emoción), luego resultan clasificadas en los tests de validación como poco verosímiles; es decir, son, en realidad, interpretaciones acústicamente incompletas. En tanto que la emoción es un fenómeno fisiológico incontrolado, cuando un actor no consigue estar objetivamente emocionado, probablemente tampoco consigue situar las condiciones de su aparato fonador exactamente del mismo modo en que lo haría si estuviese experimentando una emoción real. Entonces, efectivamente, logra que su interpretación suene como emocionada, no obstante, su construcción acústica es solo parcialmente similar a la que genera una emoción real. Eso hace su interpretación emocionada reconocible, pero poco verosímil. Es decir, alejada de la real.

Consideramos, pues, que el trabajo con informantes no actores, es decir, no entrenados en fingir estados emocionales, muy probablemente establecería una conexión perfectamente clara y directa entre la variación de las constantes fisiológicas del locutor y la codificación acústica de las emociones en la voz.

Finalmente, hemos llegado a la conclusión de que cualquier estudio que pretenda modelizar acústicamente la codificación de las emociones en el lenguaje natural, debe partir de una diferenciación muy clara entre:

a) El reconocimiento de una emoción en el sonido del habla.

b) La verosimilitud del habla emocionada.

c) La vinculación objetiva entre parámetros fisiológicos concretos y el estado emocional del hablante.

El *reconocimiento* sonoro (Cfr. RODRÍGUEZ, 1989: 201-202) de una emoción implica, simplemente, la localización por parte del oyente de algunos rasgos acústicos que permiten identificarla; este proceso no supone ninguna garantía de que la codificación de esa emoción en la voz sea completa y correcta. Igual que en un dibujo infantil reconocemos los rasgos que componen la forma de una cara viendo un circulo con unas pocas líneas elementales en su interior, nuestro oído puede reconocer el sonido de una emoción a partir de unos cuantos rasgos acústicos simples. Pero igual que reconocer la cara dibujada por un niño solo supone una conexión muy remota con la realidad, reconocer el sonido de una emoción en el habla no garantiza que lo reconocido sea algo más que un esquema demasiado elemental del sonido real del habla emocionada.

Por el contrario, hablar de *verosimilitud* del habla emocionada supone incidir en la búsqueda de garantías de similitud entre las formas sonoras que estamos escuchando y los referentes acústicos sobre la emoción que hay almacenados en la memoria de cualquier hablante. Es necesario, pues, que diferenciemos entre el fenómeno de percibir un determinado discurso oral que intenta parecer triste, y el de sentir que una voz suena realmente atravesada por la tristeza. Mientras que en el primer caso nunca calificaríamos esa tristeza como auténtica, en el segundo sí diríamos que la voz suena como verdaderamente triste. En resumen, mientras la primera no es verosímil, la segunda sí lo es. Así, debemos entender que el sonido de una emoción en el habla emocionada resulta *verosímil* cuando, además de ser reconocible, se

percibe como honesta y verdadera, es decir, como real.

Por último, hablar de *la vinculación objetiva entre parámetros fisiológicos concretos y el estado emocional del hablante* significa tomar como punto de partida el conocimiento de que toda emoción esta vinculada a fenómenos fisiológicos físicamente cuantificables, para garantizar la bondad de un corpus de habla emocionada.

7. Bibliografía citada.

AGUILAR, L, BLECUA, M, MACHUCA, M, MARIN, R.: "*Phonetic reduction processes in spontaneus speech*", Eurospeech 93, pp. 433-436, Berlín, 1993.

ARTIERES, T AND GALLINARI, P.: "*Multistate predictive Neural Networks for text-independent speaker recognition*", ESCA- EUROSPEECH, pp. 633-636, 1995.

COSNIER, J.: Psychologie des émotions et des sentiments, Retz, 1994.

BOTTE, M.C., CAN´EVET, G., DEMANY, L. ET SORIN, etc.: Psichoacoustique et Perception Auditive. Inserm-SFA-CNET. Paris, 1989.

DANTZER, R.: Las emociones, Paidós, Barcelona, 1989.

EKMAN: "*An Argument for Basic Emotions*", a Cognition and Emotion, 6, pp. , 1983

ENGSTRAD, O.: "*Sistematicy of phonetic variation in natural discourse*". Speech Comunication n. 11, pp. 337-346, North-Holland,169-200, 1992.

FARRELL, R. J. y cols.: "*Speaker recognition using Neurals Networks and conventional classifiers*" , IEEE Trans. Speech and Audio Processing, 2, pp.194-205, 1994.

FONAGY, I.: La vive voix, Payot, Paris, 1983.

GALES, M.F.J. and YOUNG, S. J.: HMM Recognition in noise using parallel model combination Proc. ESCA-EUROSPEECH, II-837-840, 1993.

GARCÍA SANCHEZ, A, MARTÍNEZ, F, RAMON, J.L.: "*Intercepstral Correlation in speaker identification/verification with very short signals*". Proc.Of ICSPAT-93. pp.1535-1540, 1993.

GARRIDO, J.M.: Modelización de patrones entonativos para la síntesis y el reconocimiento del habla. Dto de Filología española de la Universidad Autónoma de Barcelona, 1990.

GARRIDO, J.M, LLISTERRI, J, MOTA, C, RIOS, A.: "*Prosodic differences in reading style: isolated vs. contextualizated sentences*", Eurospeech 93, pp.573-576, Berlin, 1993.

GIMENEZ FERNANDEZ, A.: Marcadores emocionales en la conducta vocal. Ediciones de la Universidad Autónoma de Madrid, 1986.

IZARD, C. (ed.): Measuring Emotions in Infants and Children, Cambridge University Press, 1982.

KATAGIRI, S. and JUANG, B. H.: Discriminative feature extraction. Artificial Neural Networks for speech and vision, Ed. R. J. Mammone, 421-430, 1993.

KNAPP, N.L.: Essentials of nonverbal communication. Holt, Rinehart and Wilson, Nueva York, 1980.

LEINONEN, L., y cols.: "*Expression of emotional-motivational connotations with a one-word utterance*", JASA, 102, pp.1853-1863, 1997.

LIOU, H.S. and MAMMONE, R.J.: "*A sub-word Neural Tree Network approach to text-dependent speaker verification*" , Proc. of ICASSP, pp.357-360, 1995.

MARTINEZ, F.: Caracterización de parámetros acústicos del habla mediante señales sintetizadas. Tesis doctoral. Universidad de Murcia, 1991.

MATSUI, T. y cols.: "*Speaker recognition using HMM composition in noisy environment*", ESCA-EUROSPEECH, pp.621-624, 1995.

MATSUI, T and FURUI, S.: "*Similarity normalization method for speaker verification based on a posteriori probability*" , Proceedings ESCA WASRIV, pp.59-62, 1994.

MATSUI, T and FURUI, S.: "*Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous*". HMMs, Proceedings ICASSP, II-157-160, 1992.

MAYOR, L.: Psicologia de la emoción. Promolibro, Valencia, 1988.

MURRAY, I.R., ARNOTT, J.L.: "*Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion*". J.A.S.A n.92 (2), pp. 1097-1108, 1993.

PERROT, J. et al.: Polyphonie pour Ivan Fonagy, Ed. L'Harmattan, Paris, Montreal, 1997.

PROTOPAPAS, A, LIEBERMAN, P.: "*Fundamental frequency of phonation and perceived emotional stress*", JASA, 101, pp. 2267-2278, 1997.

REEVE, J.M.: 1994 Motivación y emoción, McGraw-Hill/Interamericana de España, S.A, Aravaca (Madrid).

RENCHLIN, M.: Psicologia. Morata, Madrid, 1987.

RODRÍGUEZ BRAVO, A.: La dimensión sonora del lenguaje audiovisual. Paidós, Barcelona, Buenos Aires, 1998

RODRIGUEZ BRAVO, A.: La construcción de una voz radiofónica. Tesis doctoral, Dto. de Comunicación Audiovisual y Publicidad de la Universidad Autónoma de Barcelona, 1989.

ROSE, R.C. y cols.: "*Integrated models of signal and bakground with application to speaker identification in noise*", IEEE Trans. Speech and Audio Processing, 2, 245-257, 1994.

ROSEMBERG y cols.: "*Sub-word unit talker verification using Hidden Markov Models*" , Proceedings ICASSP, 269-272, 1990.

SCHMIDT-ATZERT, L.: Psicología de las emociones. Herder, Barcelona, 1985.

SETLUR, A AND JACOBS, T., "*Results of a speaker verification service trial using HMM models*", ESCA-EUROSPEECH, 639-642, 1995.

SCHERER, K.R.: Facets of Emotion: Current Research. Hillsdale, London, Lawrence Erlbaum, 1988.

SCHERER, K.R y EKMAN, P.: Aproaches to Emotion, Hillsdale, London, Lawrence Erlbaum, 1984.

VROOMEN, J, COLLIER, R, MOZZICONACCI, S.: "*Duration and intonation in emotional speech*", Eurospeech 93, pp. 577-580, Berlin, 1993.