EVALUACIÓN SUBJETIVA DE UNA BASE DE DATOS DE HABLA EMOCIONAL PARA EUSKERA

Iñaki Sainz, Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Jon Sanchez, Iker Luengo, Igor Odriozola, Eneritz de Bilbao

Aholab – Dept Electrónica y telecomunicaciones. Facultad de Ingenieria.

Universidad del País Vasco, Urkijo z/g 48013 Bilbo

Email: inaki, ibon, eva, inma, ion, ikerl, igor, eneritz @aholab.ehu.es

RESUMEN

El presente artículo describe el proceso de evaluación de una base de datos de voz emocional grabada para Euskera normalizado (Euskera *batua*). El propósito de dicha evaluación no es otro que determinar la validez de la base de datos tanto para la caracterización del habla emocional como para su empleo en el desarrollo de un sistema de síntesis de habla expresiva. El corpus está formado por setecientas frases semánticamente neutras que han sido grabadas por dos actores profesionales, para 6 emociones y estilo neutro. Los resultados del test muestran que todas las emociones son correctamente identificadas muy por encima del nivel de azar, y para ambos locutores. Por lo tanto, podemos concluir que la base de datos representa un recurso lingüístico válido para los propósitos de investigación y desarrollo para los que fue originalmente diseñada.

1. INTRODUCCIÓN

Debido al progreso en las técnicas de síntesis de voz durante los últimos años, la inteligibilidad de la mayoría de los CTV (conversor de texto a voz) es prácticamente equivalente a la del habla humana. Sin embargo, la naturalidad y fluidez de las voces sintéticas está lejos de ser indistinguible de la voz natural. Una expresión emocional apropiada representa uno de los aspectos claves de la naturalidad del que aún carecen los sistemas de síntesis del habla. El habla emocional puede ser tanto una forma de reducir la monotonía de la voz sintética, como una manera de mejorar la comunicación hombre-máquina.

A lo largo de los últimos años se han llevado a cabo una serie de intentos para desarrollar un CTV expresivo [1][2][3]. El resultado final no ha alcanzado aún un nivel suficiente como para que las emociones sean percibidas como naturales. Con el fin de transmitir emociones realistas, es necesario una profunda investigación de las características prosódicas del habla emocional: contorno de pitch, duración de los fonemas y curva de energía. Y para que dicho estudio sea posible, resulta imprescindible la grabación de una buena base de datos emocional.

Si bien es cierto que las modificaciones prosódicas tienen una indudable influencia a la hora de expresar emociones [4][5], el habla expresiva resultante con estos métodos no llega a alcanzar la naturalidad deseada [6]. Y esto sucede aún cuando se utiliza directamente prosodia extraída de señales reales [7]. Ello es debido a que las emociones se expresan no solo a través de variaciones de la prosodia, sino mediante cambios en las características espectrales de la voz [8]. Dado que modelar dichas propiedades acústicas de forma explícita es realmente complejo, se pueden emplear técnicas basadas en corpus para hacerlo implícitamente.

En las soluciones basadas en corpus, cada frase está generada a partir de la concatenación de unidades extraídas de una base de datos de voz natural. El algoritmo de selección de unidades está basado en la minimización de una función coste global formada por un coste objetivo y uno de concatenación [9]. Ambos costes se hallan delimitados entre los valores 0 (la mejor elección posible) y 1 (el peor escenario posible).

El coste objetivo mide el parecido entre la unidad deseada (predicha por el módulo lingüístico y prosódico del CTV) y las unidades candidatas disponibles en la base de datos. El coste de concatenación por su parte, es calculado como una medida de la distorsión resultante al unir dos unidades candidatas. Obviamente el coste es igual a cero si ambas unidades aparecen de forma consecutiva en la base de datos.

Los sistemas de selección de unidades proporcionan buenos resultados cuando se dispone de suficientes unidades candidatas similares a la secuencia de unidades objetivo dada. De manera que no sea necesario llevar a cabo modificaciones prosódicas que supondrían la introducción de distorsiones, reduciendo así la calidad de la voz sintética. Es por ello de vital importancia contar con una base de datos de gran tamaño, poblada con una cantidad suficiente de unidades

para cada una de las emociones que se pretenda sintetizar.

A lo largo de este artículo se presenta la evaluación subjetiva de una base de datos de habla con emociones. En la sección 2 se procederá a describir el proceso de diseño y grabación de la base de datos. El protocolo de evaluación se detalla en la sección 3 y los resultados son presentados y argumentados en la sección 4. Finalmente se plasman algunas conclusiones sobre el proceso.

2. DISEÑO Y GRABACIÓN DEL CORPUS

La base de datos que se describe a continuación, fue creada teniendo en mente dos propósitos. Por una parte, se buscaba que formara el núcleo sobre el que desarrollar un sistema CTV emocional y basado en corpus para el Euskera. Por otra parte, se pretendía que fuera un recurso útil para el análisis prosódico y espectral del habla emocional.

2.1. Grabación del habla emocional

Existen diversas técnicas para grabar voz emocional, cada una con sus ventajas e inconvenientes, y entre las que podemos destacar las siguientes:

- *Emociones espontaneas*: Sin lugar a dudas las emociones más auténticas pero plantean dificultades técnicas para su grabación. Por otra parte, es prácticamente imposible controlar el contenido de las grabaciones, por lo que resulta inviable recolectar una base de datos adecuada para síntesis por corpus, debido a las restricciones en la cobertura fonética que este tipo de aplicaciones imponen.
- *Emociones provocadas*: El locutor es puesto en una situación concreta con la intención de despertar en él un estado emocional específico. Sin embargo, dado que cada persona actúa de manera diferente incluso ante un mismo estímulo, la emoción grabada mediante este método no se puede garantizar totalmente. Otra desventaja es la relativa a las consideraciones éticas que saltan a la palestra cuando es necesario evocar situaciones negativas para la recolección de emociones como pueden ser el miedo o la tristeza.
- *Emociones actuadas*: Esta técnica consiste en la lectura de un texto por parte de un actor profesional, intentado emular la emoción deseada. Este tipo de grabación es comúnmente acusada de producir emociones exageradas y faltas de naturalidad.

Finalmente, el tercer método de grabación fue el elegido por las ventajas que proporciona. Por una parte, permite controlar el contenido de la base de datos preservando la variabilidad fonética con la que fue diseñado el corpus. Y por otro lado, facilita el estudio y comparación de las características de cada emoción ya que el contenido textual se mantiene.

Durante la grabación se utilizaron textos semánticamente neutros no relacionados con las emociones. Un único corpus textual fue utilizado para la grabación de todas las emociones. La validez de esta elección fue probada experimentalmente en [10]. En lo que al contenido expresivo se refiere, se consideraron las seis “Big emotions” (Tristeza, alegría, enfado, miedo, sorpresa y asco) [11] ya que representan las más universalmente reconocibles y vinculadas con gestos faciales. Además, se realizó la grabación de estilo neutro típicamente utilizado en los sistemas CTV genéricos.

Para la grabación de aproximadamente una hora de habla para cada emoción, se seleccionaron un total de 702 frases mediante técnicas de análisis textual, garantizando tanto el balanceado fonético como la cobertura de difonemas presentes en el Euskera. El corpus fue grabado por dos actores profesionales: un actor de doblaje y una locutora de radio. Para una descripción detallada de las características del corpus consúltese [12].

3. EVALUACIÓN

Para determinar la validez del contenido emocional de la base de datos, se puso en marcha una campaña de evaluación subjetiva.

.

3.1. Diseño del test

Para descubrir si los evaluadores eran capaces de identificar la emoción expresada en las grabaciones de la base de datos, se llevó a cabo un test de elección forzada. Utilizando una interfaz basada en web se presentaron 30 estímulos de cada uno de los actores. Las señales de test se agruparon aleatoriamente de diez en diez formando formularios. Los sujetos evaluadores debían seleccionar una de las 6 posibles emociones, ya que no existía la opción de respuesta “no identificada”. Todas las frases de test eran enunciativas salvo una interrogativa. La longitud media de las señales era de 8.61 palabras, oscilando entre 4 para la más corta y 14 para la más larga.

3.1. Protocolo de evaluación

Cada uno de los evaluadores realizó el test individualmente. Las señales eran escuchadas a través de auriculares y tarjetas de sonido de gran calidad. Los oyentes no tuvieron un periodo de entrenamiento ni recibieron constancia de si sus respuestas eran correctas o no. Debían identificar las 10 señales presentadas en cada formulario, una vez hecho lo cual no se les permitía volver atrás para modificar sus contestaciones. Sin embargo, sí que podían escuchar las señales tantas veces como les fuera necesario antes de decidir la respuesta final. También se permitían descansos entre formulario y

formulario, si bien la mayoría completó el test sin necesitar ninguno.

Un total de 20 sujetos participaron en la evaluación (14 hombres y 6 mujeres) con edades que oscilaban entre 20 y 53 años. Todos los evaluadores hablaban Euskera si bien es conveniente realizar ciertos apuntes. El Euskera es una lengua minoritaria cuyo número de hablantes ha aumentado en los últimos años gracias a la promoción del idioma llevada a cabo en el sistema educativo. Podemos dividir la comunidad vasco-parlante en dos grupos: Aquellos para los que el Euskera es su primera lengua (Euskaldun Zahar) y los que lo han aprendido como segunda lengua (Euskaldun Berri). De los que completaron el test, solo 11 eran hablantes nativos.

4. RESULTADOS

Los resultados del test subjetivo se presentan en la Tabla 1. Cada una de las filas de la matriz representa la emoción expresada por el actor, mientras que las columnas muestran las emociones identificadas por los oyentes. Los valores son porcentajes y las letras simbolizan las emociones de la siguiente manera: Enfado (E), Miedo (M), Sorpresa (S), Asco (A), Alegría (L) y Tristeza (T).

La tabla también incluye los estadísticos Precisión (P) y Recall o Recuperación (R). La Precisión se calcula como el número de identificaciones correctas entre el número de respuestas asignadas a esa emoción. La Recuperación por su parte, se mide como el número de identificaciones correctas entre el número de estímulos existentes para dicha emoción.

Actores	EVALUADORES							
	E	M	S	A	L	T	P	R
E	81.5	2.5	5.5	9	-	1.5	0.78	0.82
M	0.5	64	3	-	1	31.5	0.68	0.64
S	6	2.5	73	1	17.5	-	0.80	0.73
A	15.5	4	3.5	67	2.5	-	0.86	0.67
L	0.5	0.5	5	-	94	-	0.81	0.94
T	-	20	1	0.5	1	77.5	0.66	0.78

Tabla 1. *Matriz de confusión con los resultados totales*

Puede apreciarse de manera clara que todas las emociones fueron identificadas muy por encima del nivel de azar (17%) a pesar de que el corpus estaba formado por frases semánticamente neutras que podían, a priori, dificultar el proceso de identificación. El nivel medio de reconocimiento es de 76.6%, siendo Alegría (94%) la emoción más fácilmente reconocida y Miedo (64%) la que más dificultades mostraba. Tristeza es la emoción con la menor Precisión (66%) ya que fue seleccionada como respuesta para estímulos relativos a Miedo, Asco o Enfado. La baja Recuperación pero alta Precisión para el Asco se debe al hecho de que raramente ha sido elegida como respuesta durante el test, pero cuando se ha hecho se ha acertado en la mayoría de los casos. Igualmente aunque a la inversa, Alegría ha sido una elección frecuente en el test y de ahí su alta Recuperación pero moderada Precisión.

Las Tablas 2 y 3 Ilustran las matrices de confusión para cada uno de los actores. Igual que para la matriz global, Alegría obtiene los mejores resultados en ambos casos con un 96% de identificaciones correctas para la actriz y 92% para el actor. Sin embargo, la emoción peor identificada difiere en esta ocasión, siendo Miedo (61%) para la locutora y Asco (59%) para él, aunque con una gran precisión en este último caso. La tasa de reconocimiento media es muy similar también: 75.83% para la actriz y ligeramente superior (76.50%) para el locutor masculino.

Actriz	EVALUADORES							
	E	M	S	A	L	T	P	R
E	75	-	6	15	-	3	0.76	0.75
M	1	61	4	-	1	34	0.73	0.61
S	10	2	68	-	20	-	0.82	0.68
A	13	-	2	75	1	9	0.83	0.75
L	1	-	3	-	96	-	0.81	0.96
T	-	19	-	-	1	80	0.63	0.80

Tabla 2. *Matriz de confusión para la actriz*

Actor	EVALUADORES							
	E	M	S	A	L	T	P	R
E	88	4	5	3	-	-	0.81	0.88
M	1	67	2	-	1	29	0.64	0.67
S	2	3	78	2	15	-	0.79	0.78
A	18	8	5	59	4	6	0.91	0.59
L	-	1	7	-	92	-	0.81	0.92
T	-	21	2	1	1	75	0.68	0.75

Tabla 3. *Matriz de confusión para el actor*

Volviendo a los resultados globales, las emociones que más comúnmente se confunden son Miedo y Tristeza. Miedo es identificado como Tristeza el 34% de las ocasiones, y Tristeza es identificada como Miedo el 20%. Ambas representan igualmente, el par de emociones que mayor número de veces se confunden tanto para las señales del actor como de la actriz por separado, en un rango que oscila entre el 19% y 34%. La confusión entre estas dos emociones ya se había observado por ejemplo, en la base de datos Interface para castellano [13].

Teniendo en cuenta que los evaluadores no dispusieron de una sesión de entrenamiento previo, se procedió a analizar si los resultados obtenidos en la segunda mitad del test (tasa de reconocimiento del 79.7%, rango de confianza: entre 76.26% y 83.07%) eran significativamente mejores que los de la primera (72.64%, intervalo: 69.26% - 76.06%). En esta ocasión el test t-student dictamina que la hipótesis acerca de la significancia estadística de las tasas de reconocimiento, sí resulta verdadera. Concretamente se obtiene una $t=2.85 \rightarrow p=0.0044 > 0.05$. Un análisis más detallado de los datos revela que dicha mejora del 7% en la tasa de identificación, se mantiene prácticamente constante en los distintos grupos: hombres, mujeres, actor, actriz, hablantes nativos, etc.

El resultado resulta comprensible porque durante los primeros estímulos y debido al test de respuesta forzada y a la ausencia de entrenamiento previo, es posible que los evaluadores elijan una respuesta sin estar completamente seguros. Según el test va avanzando, los oyentes aprenden la forma en la que el actor expresa determinadas emociones y ello facilita la identificación del resto por descarte.

5. CONCLUSIONES

Los resultados de la evaluación subjetiva han constatado que todas las emociones son fácilmente reconocibles para ambos locutores. Por lo tanto, esta base de datos de voz, representa un recurso lingüístico válido que permitirá tanto la caracterización del habla emocional para el Euskera, como la creación de un CTV con emociones que actualmente se encuentra en pleno desarrollo.

En lo que al diseño del test se refiere, la significativa mejora de las identificaciones durante la segunda mitad del test nos lleva a modificar la estrategia para futuras evaluaciones. Se mantendría así la estructura de elección forzada de una respuesta entre las 6 emociones posibles, añadiendo para cada señal una pregunta binaria para determinar si la identificación ha resultado sencilla o no. Esto permitiría analizar las respuestas finales con mayor eficiencia.

6. AGRADECIMIENTOS

Esta evaluación ha sido posible gracias a la financiación del Gobierno del País Vasco dentro del programa ANHITZ (ETORTEK96/114) y al MEC (TEC2006-13694-C03-02/TCM).

Los autores agradecen la colaboración de todos los sujetos que participaron en la evaluación.

7. BIBLIOGRAFÍA

[1] IIda, A., Campbell, N. Higuchi, F., & Yasumura, M. (2003). A Corpus based speech synthesis system with emotion, In *Speech Communication*, 40, pp. 161--187

[2] Murray, I.R. and Arnott, J.L. Synthesising emotions in speech: is it time to get excited?, In *ICSLP* 1996, pp. 1816--1819

[3] Bulut, Murtaza, Shrikanth S. Narayanan, & Ann K. Syrdal. Expressive speech synthesis using a concatenative synthesizer, In *ICSLP* 2002, pp. 1265--1268

[4] Vroomen, J., Collier, R., & Mozziconacci, S. J. L., Duration and Intonation in Emotional Speech, In *Eurospeech* 1993, Vol. 1, pp. 577--580

[5] Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., & Pardo, J. M., Analysis and Modeling of Emotional Speech in Spanish, In *ICPhS* 1999, pp. 957--960

[6] Schröder, M. Can emotions be synthesized without controlling voice quality?, In *Phonus 4, Research Report of the Institute of Phonetics* 1999, Saarland University, pp. 37—55

[7]Heuft, B., P ortele, T., & Rauth, M., Emotions in Time Domain Synthesis, In *ICSLP* 1996, pp.1974—1977

[8] Rank, E., & Pirker, H., Generating Emotional Speech with a Concatenative Synthesizer, In *ICSLP* 98, Vol. 3, pp. 671--674

[9] Hunt, A. and Black, A. Unit selection in a concatenative speech synthesis system using a large speech data base, In *ICASSP* 1996, pp. 373-376.Erlbaum Associates, pp. 252--262

[10] Navas, E., Hernáez, I., Luengo, I. An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS, in *IEEE Transactions on audio, speech and language processing* 2006, vol. 14, n. 4, pp. 1117--1127.

[11] Cowie, R., Cornelius, R.R. Describing the Emotional States that Are Expressed in Speech, In *Speech Communication* 2003, 40(1,2) pp. 5--32

[12] Saratxaga I, Navas E., Hernaez I., Luengo I. Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque, In *Proceedings of the LREC* 2006, pp. 2126--2129

[13] Nogueiras, A., Moreno, A., Bonafonte, A., Mariño, J.B., Speech Emotion Recognition Using Hidden Markov Models, In *Proceedings of Eurospeech* 2001, pp. 2679--2682