Reconocimiento de emociones positivas y negativas mediante el habla

1st Nerio Morán
dept. name of organization (of Aff.)
name of organization (of Aff.)
Merida, Venezuela
neriomoranp@gmail.com

2nd Jesús Perez
dept. Lasdai (of Aff.)
name of organization (of Aff.)
Mérida, Venezuela
neriomoranp@gmail.com

***Abstract*—El reconocimiento de emociones a través del habla, es una de las áreas de investigación en la interacción humano-robot con mas demanda en la actualidad. El propósito de un sistema de reconocimiento de emociones es diferenciar los estados emocionales de un hablante en las 6 emociones universales: ira, miedo, tristeza, felicidad, sorpresa y disgusto. Este sistema utiliza segmentación a largo plazo para obtener un vector de características correspondiente a cada muestra de audio. Características del dominio del tiempo y del dominio espectral son usadas entre ellas: energía, tono y los coeficientes ceptrales de Mel (MFFCs). Una base de datos propia es utilizada. Para el entrenamiento y validación del modelo es usado el algoritmo maquinas de soporte vectorial o SVM. Una comparativa con los núcleos: lineal, radial, y polinomial es mostrado. ===Resultados de los clasificadores (Para 2 clases y para las 6)====.**

***Index Terms*—Habla, Emociones, Reconocimiento, detección, Tono, energía, MFFC, SVM**

I. Introducción

A los largo de los años, ha sido el ser humano quien se ha adaptado a las diferentes formas de comunicación que ofrece una computadora. Investigaciones actuales, sugieren que la brecha de comunicación entre humanos-robots puede ser disminuida mediante el reconocimiento y adaptación de las computadores según el estado emocional de la persona [1].

El reconocimiento de emociones es realizado mediante diferentes tipo de entradas: El habla [4]–[15], imágenes de rostros [18], conductacia de la piel [16], frecuencia cardíaca [16], señales inalámbricas [17], entre otros. Dado que las señales del habla se consideran fáciles de obtener y es una de las formas de comunicación mas usadas, se le considera como una de las fuentes de información mas practicas para la clasificación de emociones [7].

Esta es la razón por la cual la clasificación de emociones mediante el habla, se ha convertido uno de los tópicos de mayor investigación en tres campos distintos: Aprendizaje de maquina, interacción humano-robot y procesamiento de señales de audio.

Las investigaciones han sugerido numerosas combinaciones de características; dentro de estas, las mas usadas para el reconocimiento de emociones son: tono, energía , MFFCs, los coeficientes dinámicos de energía de Mel (MEDC) y los formantes [4], [6]–[15].

Ademas muchas investigaciones han explorado el uso de diferentes clasificadores como : **SVM** con núcleo de función base radial **RBF-SVM** , SVM con núcleo gaussiano o **GSVM**, modelo oculto de Márkov o **HMM** , bosques aleatorios o **RDF** , potenciación del gradiente o **GB**, modelo de red neuronal de McCulloch y Pits o **MCP-NN**, perceptron multicapa o **MLP**, red neuronal probabilistica o **PNN** entre otros. El numero de clases y base de datos varia según la investigación. En el siguiente cuadro se muestra información de los trabajos relacionados:

Art.	Modelo	N°clases	% exact.	BD
[4]	GSVM	4	67.1%	Susas-DB [19]
[4]	HMM	4	70.1%	Susas-DB [19]
[4]	HMM	2	96.3%	Susas-DB [19]
[4]	GSVM	5	42.3%	Aibo-DB [20]
[15]	SVM	3	91.30%	Emo-DB [21]
[15]	SVM	3	95.09%	SJTU Chinese emotion database [15]
[6]	Rand-SVM	7	55.89%	Emo-DB [21]
[6]	RDF	7	81.05%	Emo-DB [21]
[6]	GB	7	65.23%	Emo-DB [21]
[7]	MCP NN	2	85%	Propia sin nombre
[8]	RBF-SVM	5	81%	LDC-DB [22]
[9]	RBF-SVM	5	81%	Emo-DB [21]
[10]	RBF-SVM	6	84%	Polish-DB [23]
[11]	MLP	7	83.1%	Emo-DB [21]
[11]	RDF	7	77.19%	Emo-DB [21]
[11]	PNN	7	94.1%	Emo-DB [21]
[11]	SVM	7	83.1%	Emo-DB [21]
[12]	RBF-SVM	7	86.6%	Emo-DB [21]

Figure 1. Precisión de los clasificadores

En este articulo, utilizaremos características prosódicas y espectrales con estadísticas para entrenar el clasificador SVM con diferentes núcleos y de esta manera hacer una comparativa de su desempeño para el reconocimiento de emociones.

El articulo se organiza como sigue: La segunda sección es una descripción breve de las emociones y su representación junto con las descripción de la base de datos utilizada, la tercera sección explica de manera breve los procesos involucrados en el proceso de clasificación, la cuarta sección menciona las características extraídas de las muestras de audio y la quinta sección muestra los resultados y la comparativa del clasificador SVM con los diferentes núcleos.

II. Base de datos emocional de sonido/lenguaje

Las bases de datos emocionales de audio, se pueden ser clasificadas en tres tipos según la forma en que se pide a los hablantes demostrar las emociones. [10]:

- **Lenguaje actuado**: Se les pide a los actores expresar directamente una emoción predefinida.
- **Lenguaje de la vida real**: Respuestas naturales de conversaciones, las cuales son autenticas por naturaleza.
- **Lenguaje emocional evocado**: Las emociones son inducidas y son auto-reportadas en lugar de ser etiquetadas. Es decir el hablante reconoce su propia emoción y le asigna por su mismo una etiqueta.

Entre las bases de datos que se basan en lenguaje de la vida real tenemos la base de datos: "Polish Emotional Natural Speech Database" [23] y "Automatic Classification of Emotion-Related User States in Spontaneous Children Speech" [20], basada en lenguaje actuado: "A Database of German Emotional Speech" [21] y basa en la evocación de emociones: "The enterface'05 audio-visual emotion database" [24]

A. Descripción de la base de datos

Actualmente la mayoría de bases de datos de audio disponibles para la investigación del reconocimiento de emociones, no se encuentra en español. Es por esta razón se hizo una base de datos para esta investigación. La base de datos cuenta con **X** muestras, de **X** personas, (**X** hombres y **X** mujeres). La base de datos se construyo a partir de segmentos de audio en internet. Todos las muestras de audio corresponden son en español. Para su creación fueron tomados ciertos criterios:

- Todos los segmentos de audio contienen información sobre una situación y vocalizaciones no lingüísticas.
- Cada segmento de audio tiene un tamaño entre 2-6 segundos.
- Cada segmento de audio, tiene una frecuencia de muestreo de 16Khz y se almacena en formato wav.
- El rostro de la persona que habla es tomado en cuenta para determinar la emoción correspondiente.

La base de datos contiene audios correspondientes a las 6 emociones universales: ira, miedo, felicidad, disgusto, tristeza y sorpresa descritas por Paul Ekman[2].

B. Categorías de las emociones

Las emociones se pueden clasificar desde dos puntos de vista: discretas y dimensionales. Dentro de las categorías discretas se encuentra en investigaciones de "Paul Ekman" [2], la existencia 6 emociones universales: ira, miedo, sorpresa, felicidad, tristeza, disgusto. Estas emociones se consideren universales ya que pueden ser expresadas y reconocidas por personas diferentes culturas.

El modelo dimensional conceptualiza las emociones definiendolas en 2 o 3 dimensiones. En la representación de 2 dimensiones se utiliza la valencia y excitación y fue introducida en 1980 por "James Rusell" [3]. La dimensión de la valencia se refiera a la calidad de placer de una experiencia afectiva y va desde agrado a desagrado. La dimensión de excitación se refiere a la percepción de excitación asociada a la persona con la experiencia en cuestión, y varia desde muy calmado (baja activacion) a muy emocionado (alta activacion). En la figura II-B se puede observar el modelo dimensional de Rusell con 14 emociones.

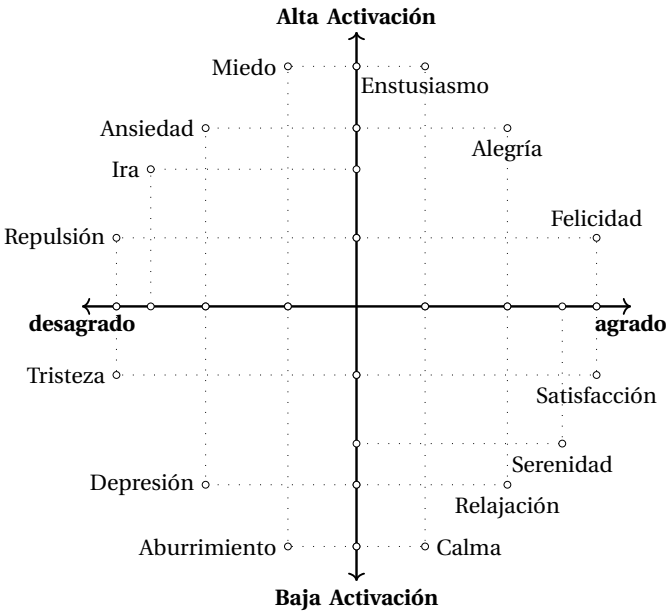

Figure 2. Modelo dimensional de Russel con 14 emociones.

III. Proceso de clasificación de las emociones mediante la voz humana

El proceso de clasificación de emociones a través del habla generalmente consta de tres pasos. El primer paso es la recolección de muestras de audio, el segundo es el proceso de extracción de características y finalmente un clasificador es creado a partir de estas características utilizando un algoritmo de aprendizaje de maquina.

A. Obtención de audio

Para la obtener el audio se usaron herramientas como youtube-dl [27]. Luego estas pistas de audio fueron transformadas y segmentadas utilizando la herramienta ffmpeg [26]. La información de los puntos de corte para cada audio fue almacenada en un archivo, que con ayuda de un script, descarga la base de datos automáticamente y almacena los segmentos de audio zen carpetas correspondientes a la emoción. Esto permite que la base de datos sea extensible, auditable y de fácil acceso.

B. Extracción de características

Los segmentos de audio se procesaron para obtener las características que serán representadas como un vector. Para realizar este proceso, un proceso de extracción a largo plazo fue llevado a cabo con la ayuda de la librería de análisis de audio "PyAudioAnalysis" [28].

El proceso de extracción de características a largo plazo, consiste en obtener el promedio de características de mediano plazo que a su vez depende del procesamiento a corto plazo de la señal. Esta forma de procesar el audio también se les conoce como segmental(corto y mediano plazo) y suprasegmental (largo-plazo) [25]. Para cada audio se utilizaron marcos de 50ms en el procesamiento a corto plazo sin solapamiento, y segmentos de 1 segundo para el procesamiento a mediano plazo sin solapamiento para finalmente obtener un vector con 64 estadísticas.

En la figura III-B se puede observar un esquema donde se muestra como se lleva a cabo el proceso de extracción de características a largo plazo.

C. Construccion del modelo de clasificación

Muchos algoritmos se clasificación han sido utilizados en diferentes investigaciones sobre el reconocimiento de emociones a través del audio. Una lista de diferentes clasificador y su desempeño puede ver en la tabla I. Todos los clasificadores necesitan datos de entrenamiento y datos de prueba, este ultimo es usado para calcular la precisión del clasificador.

El SVM es un algoritmo de aprendizaje supervisado utilizado en muchas aplicaciones de reconocimiento de patrones. Este algoritmo tiene la característica de que incluso cuando es entrenado con pocos datos provee buenos resultados [4], [6], [8]–[12], [15]. El proceso de clasificación utilizara los vectores provenientes del modulo de extracción de características para su entrenamiento y validación. En la figura III-C se puede observar el diagrama del modulo

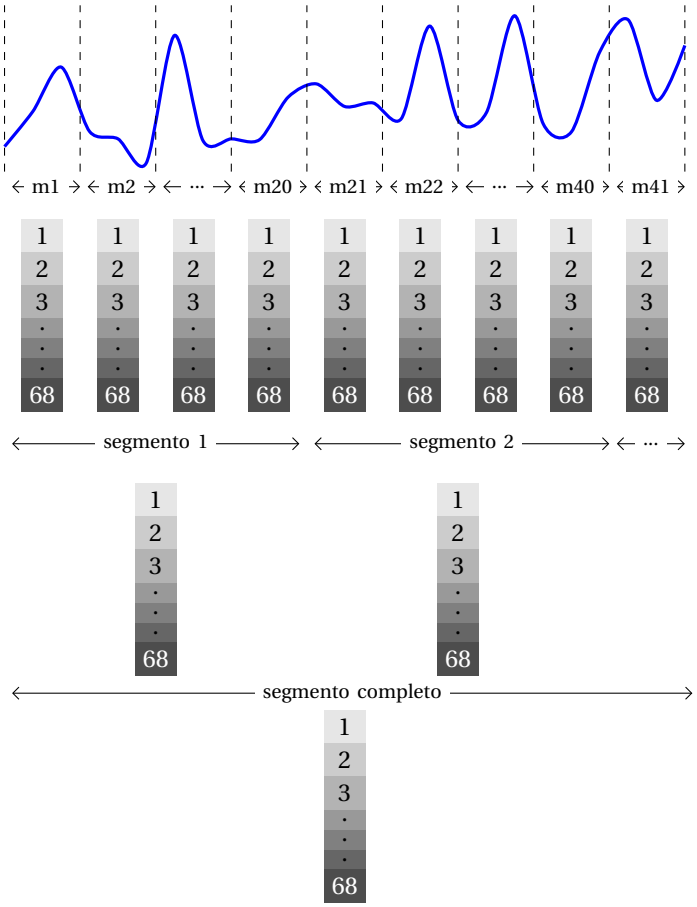

Figure 3. Procesamiento a largo plazo o suprasegmental del audio.

del proceso de clasificación similar a [10]. En este articulo se utilizara un clasificador SVM con 3 diferentes núcleos: lineal, radial y polinomial. La implementación es basada en la librería scikit-learn [29], para cada núcleo los mejores hiper-parametros serán seleccionados.

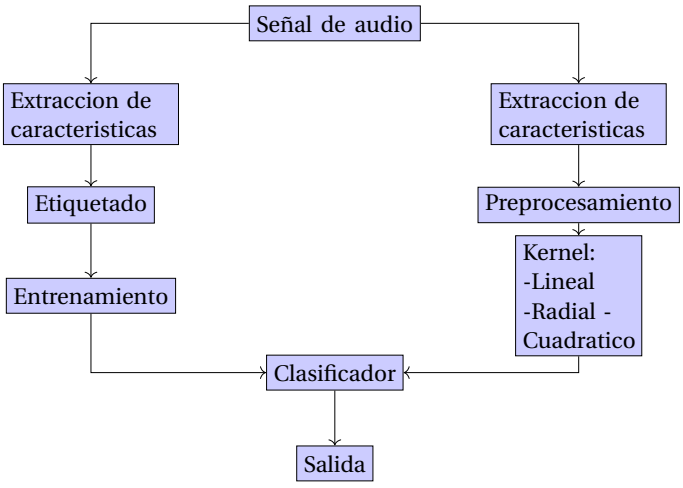

Figure 4. Diagrama del modulo del proceso de clasificación

IV. Características del sonido/habla

Las señales de audio, y en particular aquellas que contienen contenido emocional, se caracterizan por tener un gran numero de información que refleja características emocionales. Una de las cosas mas importantes en las investigaciones de reconocimiento de emociones a través del habla, es seleccionar las características adecuadas para reconocer un mayor numero de emociones.

Para esta investigación se utilizaron 3 tipos de características: Características del dominio del tiempo, del dominio del frecuencia y del dominio ceptral. Todas las características fueron obtenidas mediante la librería PyaudioAnalysis [28], una descripción formal de las características junto con sus algoritmos puede encontrarse en [31]. A continuación se muestra una descripción de la características que se seleccionaron para esta investigación.

1) **Taza de cruces por cero (ZRC)**: Es la taza de cambios de signo a lo largo de una señal, por ejemplo la taza en la cual cambia de positivo a negativo o viceversa. Esta característica es ampliamente utilizada en aplicaciones de reconocimiento del habla,reconocimiento de emociones y recuperación de información musical. Dentro de las investigaciones que utilizan esta característica tenemos [10].
2) **Energía o Potencia de la señal**: La energía se define como la suma de los cuadrados de las muestras, usualmente se normaliza dividendo entre la longitud de la muestra. La energía es la característica mas básica en el procesamiento de señales del habla. Este juega un papel importante en el reconocimiento de emociones. Por ejemplo, las emociones como la felicidad o la Ira contienen una mayor energía en comparación a la tristeza. La mayoría de las investigaciones utilizan esta característica [4], [6], [11], [12], [14], [15].
3) **Entropia de la energía**: La entropia de una señal, es una medida de la cantidad de información que esta trae, su excelente desempeño en aplicaciones de reconocimiento de voz y actividad de señal ha generado que esta característica sea usada incluso en el contexto del procesamiento de imágenes. También es interpretada como una medida de cambios abruptos en el nivel de energía de una señal de audio. Es usada en el reconocimiento de voz, detección de actividad, análisis espectral y clasificación de marcos (separación de ruido y tono).
4) **Centroide Espectral**: Es el centro de gravedad del espectro, denota el punto de balance la magnitud del espectro. Este captura las características concernientes a la pendiente espectral.
5) **Tono**: El tono estrictamente hablando es la frecuencia percibida, pero en la literatura el tono y la frecuencia fundamental son términos intercambiables. Para la señales periódicas, la frecuencia equivalente a la longitud del periodo fundamental de la señal es conocido como la frecuencia fundamental [31]. Es una característica fundamental utilizada dentro de las investigaciones: [4]–[8], [10]–[12], [14], [15].
6) **La entropía espectral**: Se calcula de manera similar a la entropía de energía, sin embargo, se aplica en el dominio de la frecuencia. La entropía espectral es una característica utilizada en la detección de actividad del habla y reconocimiento. Ademas, ha mostrado mejorar la precisión del reconocimiento debido a que es menos afectada por el ruido cuando es usada como característica adicional [30].
7) **Flujo espectral**: Denota el cambio del espectro local de una señal emocional. El flujo espectral indica que tan rápido cambia el espectro de potencia dentro de los marcos en la señal. Es usada en las investigaciones [12], [13].
8) **Caída espectral**: La energía de una señal de habla que contiene información de una emoción es encontrada dentro de determinado rango de frecuencias, La caída espectral indica el contenido de las frecuencias por debajo de las cuales ciertas fracciones de la cantidad total de energía se mantienen. Dependiendo de las características de la señal, es utilizado un valor entre 0.95 y 0.85 para el calculo. La caída espectral ademas provee información de la forma espectral con al cual se determina la cantidad de componentes de frecuencias altas disponibles en la señal. Es usado en la investigación [12], [13].
9) **MFFCs**: Los coeficientes ceptrales de las frecuencias de Mel han sido muy popular en el campo del análisis de voz. En la práctica, los MFCC son los coeficientes discretos de la transformación coseno del espectro de potencia logarítmica en la escala de Mel. Los MFCCs ha sido ampliamente utilizado en reconocimiento de voz, agrupamiento de altavoces, reconocimiento de emociones y muchos otros tipos de aplicaciones de análisis de audio y aprendizaje de maquina. Caracterizan la magnitud del espectro y por lo general son usados los 12 primeros coeficientes. En la gran mayoría de investigaciones los MFFCs han mostrado ser la característica que mejores cualidades tiene para el reconocimiento de emociones [4], [6], [8]–[12], [14], [15].
10) **Vector cromático** : Esta es una representación en 12 dimensiones de la energía espectral de una señal de audio. Este es un descriptor ampliamente utilizado, principalmente en aplicaciones relacionadas con la música, sin embargo, también se ha utilizado en el análisis de voz. El vector de croma se calcula agrupando los coeficientes espectrales de un marco en 12 contenedores representando las 12 clases de tono moderado de la música de tipo occidental.
11) **Desviación estándar del vector cromático**: Se añade la desviación estándar el vector cromático.

En la figura IV, se puede observar el modulo de extracción de características.

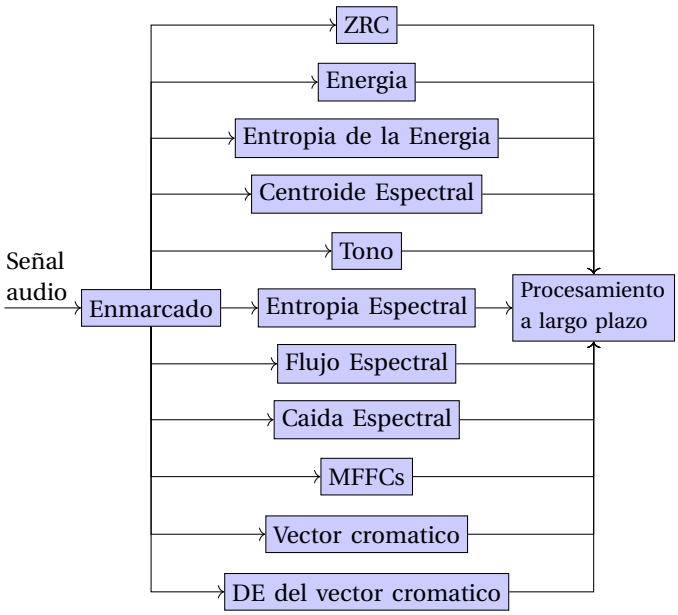

Figure 5. Diagrama del modulo de extracción de características

En total se genera un vector de 34 características por cada marco, que sera usado para generar un vector de 68, cuyos elementos corresponden al promedio y desviación estándar de las 34 características obtenidas mediante el procesamiento a largo plazo.

V. Experimento y resultados

- Matrices de confunsion

VI. Conclusiones

References

[1] R. W. Picard, "Toward computers that recognize and respond to user emotion," in IBM Systems Journal, vol. 39, no. 3.4, pp. 705-719, 2000. doi: 10.1147/sj.393.0705

[2] Ekman Paul & V. Friesen Wallace. (1971). Constants across cultures in the face and emotion. Journal of personality and social psychology. 17. 124-9. 10.1037/h0030377.

[3] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161-1178.

[4] Kwon Oh-Wook & Chan Kwokleung & Hao Jiucang & Lee Te-Won. (2003). Emotion recognition by speech signals. Proc Eurospeech.

[5] Sengupta, S. (2015). An Emotion Based Speech Analysis.

[6] M. Ghai, S. Lal, S. Duggal and S. Manik, "Emotion recognition on speech signals using machine learning," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala, 2017, pp. 34-39.

[7] Kirandzhiska, V., & Ackovska, N. (2012). Sound features used in emotion classification.

[8] Vijayalakshmi P., Anny Leemam A. (2014). Real-time Speech Emotion Recognition Using. Support Vector Machine.

[9] Chavhan, Y. D., Yelure, B. S., & Tayade, K. N. (2015, February). Speech emotion recognition using RBF kernel of LIBSVM. In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on (pp. 1132-1135). IEEE.

[10] Dahake, P. P., Shaw, K., & Malathi, P. (2016, September). Speaker dependent speech emotion recognition using MFCC and Support Vector Machine. In Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on (pp. 1080-1084). IEEE.

[11] Iliou, T., & Anagnostopoulos, C. N. (2010). Classification on speech emotion recognition-a comparative study. animation, 4, 5.

[12] Chandrasekar, P., Chapaneri, S., & Jayaswal, D. (2014). Emotion Recognition from Speech using Discriminative Features.

[13] Palo Hemanta & Kumar Pawan & Mohanty Narayan. (2017). Emotional Speech Recognition using Optimized Features. INTERNATIONAL JOURNAL OF RESEARCH IN ELECTRONICS AND COMPUTER ENGINEERING. VOL. 5. 4-9.

[14] Sinith, M. S., Aswathi, E., Deepa, T. M., Shameema, C. P., & Rajan, S. (2015, December). Emotion recognition from audio signals using Support Vector Machine. In Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in (pp. 139-144). IEEE.

[15] Pan, Y., Shen, P., & Shen, L. (2012). Feature extraction and selection in speech emotion recognition. Proceeding of the onlinepresent. org, 2, 64-69.

[16] Ménard, M., Richard, P., Hamdi, H., Daucé, B., & Yamaguchi, T. (2015, February). Emotion Recognition based on Heart Rate and Skin Conductance. In PhyCS (pp. 26-32).

[17] Zhao, M., Adib, F., & Katabi, D. (2016, October). Emotion recognition using wireless signals. In Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking (pp. 95-108). ACM.

[18] Hsu, S. C., Huang, H. H., & Huang, C. L. (2017, April). Facial expression recognition for human-robot interaction. In Robotic Computing (IRC), IEEE International Conference on (pp. 1-7). IEEE.

[19] J. H. L. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting Started with the SUSAS: Speech Under Simulated and Actual Stress Database," Robust Speech Processing Laboratory April 15, 1998.

[20] Steidl, S. (2009). Automatic classification of emotion related user states in spontaneous children's speech (pp. 1-250). Erlangen, Germany: University of Erlangen-Nuremberg.

[21] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In Ninth European Conference on Speech Communication and Technology.

[22] https://www.ldc.upenn.edu/

[23] Kamińska Dorota & Sapiński Tomasz & Pelikant Adam. (2015). Polish Emotional Natural Speech Database.

[24] Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006, April). The enterface'05 audio-visual emotion database. In Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on (pp. 8-8). IEEE.

[25] Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review, 43(2), 155-177.

[26] Bellard, F., & Niedermayer, M. (2012). FFmpeg. Availab from: http://ffmpeg.org, 3.
[27] Gonzalez, R. G. (2006). YouTube-dl: download videos from youtube.com
[28] Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. PloS one, 10(12), e0144610.
[29] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.
[30] Toh, A. M., Togneri, R., & Nordholm, S. (2005). Spectral entropy as speech features for speech recognition. Proceedings of PEECS, 1.
[31] Giannakopoulos, T., & Pikrakis, A. (2014). Introduction to audio analysis: a MATLAB® approach. Academic Press.