XIII Congresso Ibérico de Acústica
55º Congreso Español de Acústica
TECNIACÚSTICA 2024
Faro, 11 a 13 setembro 2024

EXPLORANDO LA PERSONALIZACIÓN DE MODELOS EN EL RECONOCIMIENTO DE EMOCIONES EN EL HABLA

Carlos Castorena[1], Jesús López-Ballester[1], Maximo Cobos[1], Francesc J. Ferri[1]

[1]Universitat de València, España
{carlos.castorena@uv.es / jesus.lopez-ballester@uv.es / maximo.cobos@uv.es / francesc.ferri@uv.es}

Resumen

El reconocimiento preciso de emociones en el habla es fundamental en numerosas aplicaciones, desde la salud mental hasta la interacción humano-máquina. Aunque los modelos de aprendizaje profundo han demostrado ser eficaces en esta tarea, su desempeño puede verse comprometido al enfrentarse a nuevos sujetos de prueba, incluso cuando se basan en modelos pre entrenados con una gran diversidad de datos. En este trabajo, investigamos la importancia de la personalización de modelos para el reconocimiento de emociones en el habla. Observamos que los modelos generales, a pesar de su buen rendimiento original, tienden a mostrar una disminución en su capacidad predictiva al enfrentarse a sujetos nuevos. Bajo este contexto, presentamos un estudio que explora cómo la personalización con pocas muestras de un nuevo sujeto puede mejorar significativamente el rendimiento del modelo pre entrenado utilizando técnicas de ajuste fino y transferencia de aprendizaje. Además, discutimos la relevancia de estos hallazgos en contextos prácticos y destacamos las implicaciones para futuras investigaciones en el campo de la detección de emociones mediante IA.

Palabras clave: Reconocimiento de emociones, Deep Learning

Abstract

The accurate speech emotion recognition is crucial in numerous applications, from mental health to human-machine interaction. Although deep learning models have proven effective in this task, their performance may be compromised when faced with new test subjects, even when based on pretrained models with a wide variety of data. In this study, we investigate the importance of model personalization for speech emotion recognition. We observed that general models, despite their good original performance, tend to show a decrease in predictive capacity when encountering new subjects. Within this context, we present a study that explores how personalization with a few samples from a new subject can significantly improve the performance of the pretrained model using fine-tuning and transfer learning techniques. Furthermore, we discuss the relevance of these findings in practical contexts and highlight implications for future research in the field of emotion detection through AI.

Keywords: Emotion recognition, Deep Learning

PACS nº. 43.72.Bs

1 Introducción

El Reconocimiento de Emociones en el Habla (SER, por sus siglas en inglés) es un desafío crucial en el procesamiento del lenguaje natural y la interacción humano-máquina, cuyo objetivo es identificar diversas emociones utilizando únicamente señales de voz. Aunque los sistemas actuales pueden procesar el habla de manera eficiente para tareas como transcripción o traducción [1], su desempeño disminuye considerablemente en tareas de SER debido a múltiples desafíos inherentes. Estos desafíos incluyen la falta de robustez del modelo al tratar con sujetos desconocidos, la dificultad de modelar y caracterizar las emociones en el habla, y las dependencias semánticas que algunos modelos desarrollan durante el entrenamiento.

El SER tiene múltiples y diversas aplicaciones en la vida cotidiana. Por ejemplo, podría implementarse en el servicio al cliente para monitorear y mejorar las interacciones detectando frustración o insatisfacción [2], y en el ámbito de la salud para el monitoreo de la salud mental [3], proporcionando datos valiosos para el diagnóstico y manejo de condiciones como la tristeza y el miedo. Además, el SER puede aplicarse en sistemas automotrices para interpretar el estado emocional del conductor, siendo este último escenario de particular interés para el desarrollo de este trabajo.

Generalmente, cuando un modelo de SER se entrena y se prueba con el mismo hablante, se conoce como SER dependiente del hablante [4]. Este enfoque, aunque ofrece alta precisión (alrededor del 90% en ACC), está limitado a aplicaciones específicas debido a su falta de generalización. En contraste, el SER independiente del hablante es una tarea mucho más exigente, ya que reconoce que diferentes hablantes muestran variaciones en los componentes del habla al expresar la misma emoción, lo que lleva a cambios en la distribución de las características del habla emocional entre diferentes hablantes [4], haciendo necesario ajustar los modelos en función del hablante.

En este trabajo analizamos cómo, con un número reducido de muestras correctamente etiquetadas, podemos ajustar el modelo preentrenado para proporcionar mejores predicciones de un nuevo hablante. Primero, analizamos el comportamiento basado en la selección aleatoria de muestras y luego discutimos el rendimiento considerando muestras clave que proporcionan un resultado de clasificación notablemente mas alto que el resto de muestras. Existen trabajos relacionados a esto [5], sin embargo, no llevan al limite las capacidades del clasificador usando un numero reducido de muestras, analizando únicamente diferentes tamaños del conjunto de Entrenamiento.

2 Metodología

2.1 Conjunto de datos

Existen diversos conjuntos de datos relacionados con la detección de emociones aplicados al contexto del habla, que varían en cuanto al número de sujetos, idiomas representados, cantidad y tipos de emociones, y número de muestras, entre otros aspectos. En este trabajo, hemos empleado varios conjuntos de datos ampliamente utilizados en la investigación de SER: el MSP-Podcast [6], que consiste en podcasts en inglés con etiquetas de valencia, activación y dominancia, incluyendo 83 horas de grabaciones de 60 hablantes y más de 50,000 muestras; el IEMOCAP [7], con 12 horas de diálogos de 10 hablantes y un total de 10,039 muestras; el EMO-DB [8] de Berlín, con grabaciones de 10 actores expresando 7 emociones, sumando 800 muestras; el MESD [9] en español, con grabaciones de 3 hablantes y 864 muestras en total; el RAVDESS [10], con 24 actores nativos en inglés y 7356 grabaciones; y el ESD [11], con 350 diálogos por emoción expresados por hablantes nativos de inglés y

mandarín. Para estandarizar los criterios, en los conjuntos de datos con etiquetas categóricas, solo se consideran cinco emociones (Felicidad, Enojo, Tristeza, Sorpresa y Neutral), descartando las demás.

2.2 Modelo base

Nuestro método se basa en un modelo profundo de clasificación preentrenado con diversos hablantes mostrado en la Figura 1. Primero, utilizamos una adaptación de Wav2Vec, que fue entrenado y validado con MSP-Podcast e IEMOCAP [12], respectivamente, para obtener una representación profunda de los audios. Estas capas, a las cual nos referiremos como extractor de características, permanecerán congeladas a lo largo de los experimentos. Posteriormente, esta representación pasa por un clasificador denso de 3 capas con 64, 32 y 5 neuronas, utilizando funciones de activación ReLU para las dos primeras capas y Softmax para la última capa, con el fin de proporcionar predicciones categóricas. Para el entrenamiento de las capas de clasificación, utilizamos las bases de datos EMO-DB, MESD y RAVDESS, considerando los 37 hablantes, entrenando durante 200 épocas con una función de pérdida de Entropía Cruzada Categórica en lotes de 64 muestras.

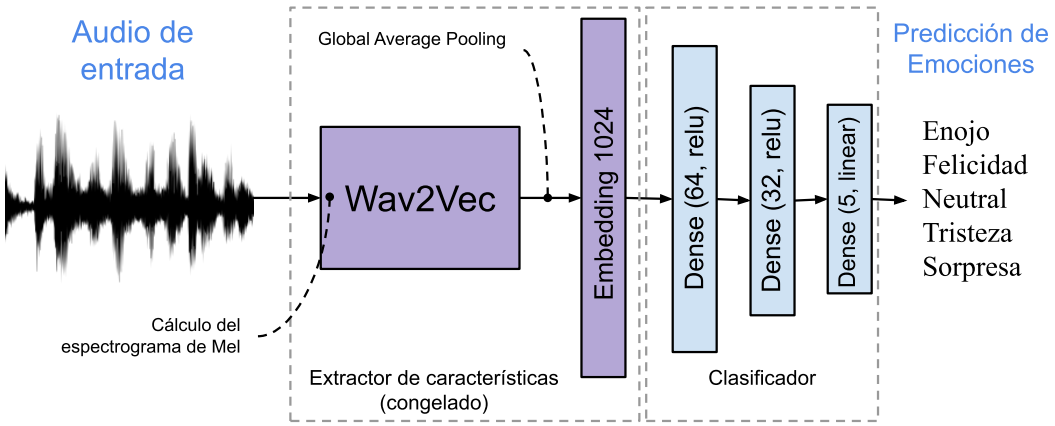

Figura 1 - Modelo base preentenado para la clasificacion de 5 emociones.

2.3 Estrategia para nuevos hablantes

Nuestra estrategia se centra en adaptar los pesos del modelo base para nuevos hablantes, extraídos de la base de datos ESD. Seleccionamos a 3 hablantes nativos de inglés y 3 nativos de mandarín, particionando los datos de cada hablante en dos conjuntos: Entrenamiento y Prueba, en una proporción de 70-30. La partición de Entrenamiento incluye todas las muestras disponibles para la selección (sin que necesariamente se utilicen todas), mientras que la partición de Prueba se mantiene constante en todos los experimentos del mismo hablante.

Incrementalmente, seleccionamos 5 muestras aleatorias en cada paso, representando las 5 emociones. Estas muestras, junto con el acumulado de los pasos anteriores, se utilizan para ajustar el modelo base durante 30 épocas, permitiendo modificar únicamente los pesos de las capas de clasificación. Este proceso da como resultado un modelo reentrenado a lo largo de 150 épocas, con un número de muestras creciente entre 5 y 25, completamente balanceado y utilizando la misma configuración de función de pérdida que el modelo base. En cada paso, utilizamos la métrica ACC para medir el rendimiento del método sobre la partición de Prueba, que se mantiene balanceada en sus 5 categorías. Cada uno de estos experimentos se repite 5 veces para cada nuevo hablante y lo hemos denominado *Aleatorio*.

Adicionalmente, presentamos una estrategia (a la cual nos referiremos con el nombre de *Mejores*) que selecciona las muestras más representativas en cada paso. Se seleccionan muestras aleatorias, se reentrena el modelo y se mide el rendimiento 200 veces en cada paso, manteniendo únicamente las 5 muestras que alcanzan el ACC más alto en la partición de Entrenamiento y añadiendo las mejores muestras encontradas en pasos anteriores. El ajuste del modelo base en la estrategia *Mejores*, siempre se hace con los mismos parámetros que en el experimento *Aleatorio.* Para facilitar la presentación de estas 200 repeticiones en cada paso, hacemos uso de un diagrama de violín que muestra la dispersión de los ACC de la partición de Prueba de todas las repeticiones.

3 Resultados y discusión

Mostramos los resultados obtenidos para los 6 nuevos hablantes en la Figura 2. En ella, se presentan las 5 repeticiones (líneas grises) y su respectivo promedio (línea negra) de la estrategia Aleatoria. Además, se muestran las dispersiones de las 200 repeticiones (violines azules), el promedio (línea continua azul) y el extremo superior (línea azul punteada) para la estrategia Mejores. También se ha marcado una línea de referencia que indica el ACC usando el modelo base sin ajustar, y por último, la línea Máximo, que representa el ACC alcanzado al usar toda la partición de Entrenamiento.

En todos los casos, observamos que existe un margen de mejora, representado por la diferencia entre las líneas Referencia y Máximo, indicando que el ACC mejora cuando el modelo base es ajustado con datos del nuevo hablante. Otro punto a destacar es que, en casi todos los hablantes, la estrategia Mejores alcanza un valor similar al Máximo utilizando únicamente entre 10 y 15 muestras. Esto sugiere que hay una configuración específica de muestras que equivale a usar todo el conjunto de Entrenamiento.

Después del primer paso, habiendo entrenado con 5 muestras, los resultados promedio de ACC para ambas estrategias son similares, ya que no hay diferencia significativa entre ellas en este punto. Esto se debe a que ambas parten del modelo base y la probabilidad de seleccionar las muestras iniciales es similar. Por lo tanto, con un número suficiente de repeticiones en la estrategia Aleatoria, también se podría alcanzar la configuración de muestras que proporciona el mejor ACC. Conforme el número de pasos se incrementa, la diferencia entre las estrategias crece, ya que el nuevo ajuste para la estrategia Mejores se basa en el mejor estado posible del modelo y no en el más probable, a diferencia de la estrategia Aleatoria. Además, aunque en un paso se pueda partir de un buen ACC, siempre existe la posibilidad de empeorar el resultado, aunque esta probabilidad disminuye a medida que el número de pasos aumenta.

En un escenario real, seleccionar las mejores muestras no es posible ya que se necesitarían conocer las etiquetas reales de antemano. Sin embargo, este experimento revela que hay un conjunto reducido de muestras que logra la tarea de manera satisfactoria. Esto implica que, identificando las características que hacen que una muestra sea efectiva, se podría solicitar la etiqueta solo de esas muestras, reduciendo al mínimo la necesidad de etiquetado por parte del nuevo usuario.

Finalmente, aunque la estrategia Aleatoria siempre rinde por debajo de la estrategia Mejores, también resulta evidente que hay una mejora respecto al estado inicial, por lo que sigue siendo una alternativa viable para mejorar el rendimiento del clasificador.

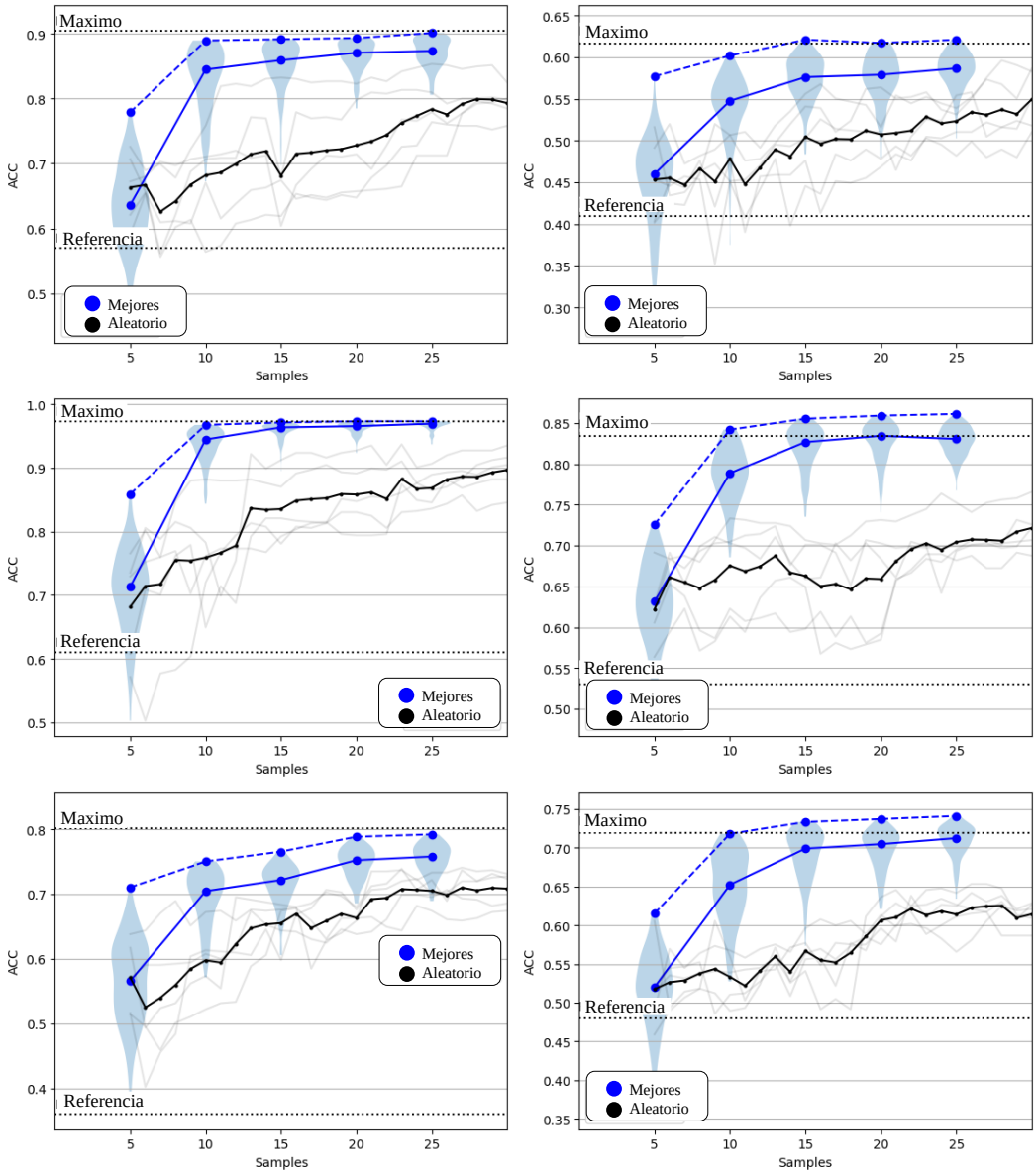

Figura 2 - Resultados de ACC para 6 diferentes nuevos hablantes usando un maximo de 25 muestras.

4 Conclusiones y Trabajo futuro

Este trabajo ha abordado el desafío del Reconocimiento de Emociones en el Habla (SER) al adaptar un modelo preentrenado a nuevos hablantes utilizando un número reducido de muestras. A través de nuestra metodología, que incluye tanto una estrategia aleatoria como una optimizada para la selección de muestras, demostramos que es posible mejorar significativamente la precisión del modelo base. Además, se ha mostrado que, incluso en escenarios donde no se puede predecir qué muestras serán las más efectivas, la estrategia aleatoria aún puede ofrecer mejoras significativas. Este hallazgo es crucial para aplicaciones prácticas donde el etiquetado exhaustivo no es factible.

Este estudio no solo reafirma la viabilidad de ajustar modelos SER a nuevos hablantes con datos limitados, sino que también destaca la potencial eficiencia de estrategias de selección de muestras más

inteligentes. Esto abre la puerta a futuros desarrollos en el campo del SER, donde la personalización del modelo puede lograrse de manera más rápida y con menos recursos, mejorando así la interacción humano-máquina en diversas aplicaciones prácticas.

Agradecimientos

Agradecemos a la Agencia Española de investigación (AEI) y al European Regional Development Fund (ERDF) por financiar parcialmente esta investigación por medio de los proyectos TED2021-131003B-C21 and PID2022-137048OB-C41financiados por MCIN/AEI/10.13039/501100011033 y por la "EU Union NextGenerationEU/PRTR". Además, agradecemos a la Generalitat Valenciana que financia este trabajo por medio del programa Santiago Grisolia (GRISOLIAP/2021/060, CPI-21-232).

Referencias

[1] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. Mcleavey, & I. Sutskever. (2023). "Robust speech recognition via large-scale weak supervision". Proceedings of the 40th International Conference on Machine Learning, vol. 202, pp. 28 492–28 518.

[2] X. Li & R. Lin. (2021). "Speech emotion recognition for power customer service". 7th International Conference on Computer and Communications (ICCC), pp. 514–518.

[3] N. Elsayed, Z. ElSayed, N. Asadizanjani, M. Ozer, A. Abdelgawad, & M. Bayoumi. (2022). "Speech emotion recognition using supervised deep recurrent system for mental health monitoring". IEEE 8th World Forum on Internet of Things (WF-IoT), pp. 1–6.

[4] M. Abdelwahab & C. Busso. (2017). "Incremental adaptation using active learning for acoustic emotion recognition". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5160–5164.

[5] M. Abdelwahab & C. Busso, (2015) "Supervised domain adaptation for emotion recognition from speech". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5058–5062.

[6] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, & S. S. Narayanan. (2008). "Iemocap: Interactive emotional dyadic motion capture database". Language Resources and Evaluation, vol. 42, pp. 335–359.

[7] R. Lotfian & C. Busso. (2019). "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings". IEEE Transactions on Affective Computing, vol. 10, pp. 471–483.

[8] F. Burkhardt, W. F. Sendlmeier, F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, & B. Weiss. (2005). "A database of german emotional speech see profile a database of german emotional speech". Proc. Interspeech 2005, pp. 1517–1520.

[9] M. M. Duville, L. M. Alonso-Valerdi, & D. I. Ibarra-Zarate. (2021). "Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody". Data 2021, Vol. 6, vol. 6, p. 130.

[10] S. R. Livingstone & F. A. Russo. (2018) "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english". PLOS ONE, vol. 13, p. e0196391.

[11] K. Zhou, B. Sisman, R. Liu, & H. Li. (2021). "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset". ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2021-June, pp. 920–924.

[12] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, & B. W. Schuller. (2023). "Dawn of the transformer era in speech emotion recognition: Closing the valence gap". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, pp. 10 745–10 759.