Detección automática de emociones a partir de la voz combinando bases de datos para aumentar el entrenamiento

Miguel Ángel Pastor Yoldi, Dayana Ribas González, Alfonso Ortega Giménez

Voice input Voice output Laboratory (ViVoLab)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: 738470@unizar.es

Resumen

La voz es la vía de comunicación más natural para el ser humano, aportando tanto información lingüística, como del estado emocional del hablante. Con el objetivo de aumentar la precisión de los sistemas de reconocimiento de emociones en voz, combinamos 3 datasets en el entrenamiento, logrando aumentar su precisión.

Introducción

Cada vez es más común la interacción humano-máquina, tanto en el ámbito profesional como en el doméstico. En los últimos años han surgido algunas aplicaciones capaces de llevar a cabo diversas tareas mediante comandos orales. Sin embargo, y pese al gran desarrollo de las tecnologías de reconocimiento automático del habla, siguen ofreciendo precisiones insuficientes en un aspecto fundamental de la comunicación, como es el reconocimiento de emociones [1].

Uno de los problemas para desarrollar sistemas automáticos de reconocimiento de emociones, es el reducido número de audios y locutores que contienen las bases de datos de habla emocionada. Esto limita la capacidad de generalización de los sistemas desarrollados. Para tratar de mitigar este problema e impactar en el desempeño del sistema, en este trabajo proponemos la combinación de diversas bases de datos para aumentar el set de entrenamiento. Para evaluar la hipótesis implementamos un sistema básico de detección de emociones que se basa en representar la información paralingüística de la señal de voz y clasificarla con un modelo estadístico [2]. En este caso utilizaremos el modelo SVM (en inglés: *Support Vector Machine)* que tiene múltiples antecedentes positivos para esta tarea [3].

Marco Experimental

Para este experimento, las bases de datos seleccionadas han sido las siguientes: EmoDb [4], RAVDESS [5] e IEMOCAP [6].

Dataset	Duración	Idioma	Locutores
EmoDb	45 min	Alemán	10
RAVDESS	3 horas	Inglés	24
IEMOCAP	12 horas	Inglés	10

Las emociones de cada dataset han sido reducidas a 4 (neutralidad, alegría, ira y tristeza), ya que el resto de emociones, o no estaban incluidas en todos los datasets, o estaban muy poco representadas en ellos. Este es un problema común en los datasets de emociones, en los que las muestras de emociones como ‘sorpresa’ o ‘asco’ son bastante escasas, mientras que las de neutralidad son mucho más abundantes.

Como extractores de características, hemos optado por GeMAPS, eGeMAPS [7] y ComParE [8], que extraen de los audios parámetros de energía espectral, frecuencia fundamental y espectro de la señal y sobre estos calculan diversos funcionales estadísticos como la media, varianza, curtosis, percentiles, etc. El más completo de ellos es ComParE, con 6373 parámetros en total (véase la configuración exacta de los parámetros y funcionales en las tablas 2 y 3 de [8]).

Para llevar a cabo la validación cruzada, dividimos cada base de datos en 5 grupos, cada uno con el mismo número de locutores, y con el mismo número de hombres que de mujeres. De esta forma evitamos que el sistema aprenda únicamente a reconocer la voz de los hablantes incluidos en el datasets, forzándola a generalizar.

En cuanto al SVM, tras unas primeras pruebas, se comprobó que el kernel gaussiano era el que mejores prestaciones ofrecía por un margen considerable y por tanto fue el kernel finalmente seleccionado. Se procedió a realizar un barrido exponencial de los parámetros C y γ, para los que obtuvimos unos resultados prácticamente idénticos en todos los experimentos realizados.

Para medir el desempeño del sistema se emplea la métrica UAR (en inglés: *Unweighted Average Recall*) (1).

$$UAR = 0.5 \frac{TP}{TP+FN} + 0.5 \frac{TN}{FP+TN} \quad (1)$$

Esta medida realiza una suma no ponderada de la precisión en la detección de cada una de las emociones. Para ello utiliza la cantidad de verdaderos y falsos positivos y negativos (TP, FP, TN, FN) en la matriz de confusión. Esta medida se emplea comúnmente en el reconocimiento de emociones para evitar perjudicar a las emociones menos representadas.

Resultados

Los resultados obtenidos se muestran en la Tabla 1.

Tabla 1. Resultados de UAR para la detección de emociones usando EmoDb, RAVDESS, IEMOCAP.

Train	EmoDb	RAVDESS	IEMOCAP	Todos
Dataset para Evaluación: EmoDb				
GeMAPS	78,21%	67,67%	74,35%	77,40%
EGeMAPS	78,94%	69,04%	73,48%	79,67%
Compare	82,50%	61,33%	70,49%	**83,53%**
Dataset para Evaluación: RAVDESS				
GeMAPS	54,04%	60,43%	56,12%	64,06%
EGeMAPS	53,52%	61,28%	56,51%	63,91%
Compare	51,04%	65,60%	52,99%	**67,81%**
Dataset para Evaluación: IEMOCAP				
GeMAPS	44,77%	43,17%	58,91%	58,52%
EGeMAPS	43,51%	43,75%	59,58%	59,79%
Compare	44,71%	46,70%	62,79%	**63,54%**

En términos comparativos, se puede observar que, para todos los datasets, obtenemos una mejora en los resultados cuando combinamos los 3 datasets en el ajuste del SVM. Estas mejoras son bastante pequeñas (del 1% para EmoDb e IEMOCAP, y un 2% para RAVDESS), pero interesantes considerando que las emociones en cada datasets están realizadas en diferentes idiomas, por personas diferentes, o sea que hay mucha variabilidad en estas clases.

También observamos en los resultados que, aun no incluyendo ningún audio del dataset de test en el entrenamiento, obtenemos unos resultados que no distan demasiado de los obtenidos al entrenar con la propia base de datos. Es decir, que, a pesar de la variabilidad introducida por el aumento de datos, no se perjudica el desempeño del sistema. Véase este efecto especialmente en EmoDb, que es significativamente más reducida que las otras dos bases de datos. Aunque también es importante observar que este dataset está en alemán, idioma que no aparece en ninguna de las otras dos bases de datos.

Conclusiones y líneas futuras

Hemos visto que la combinación de datasets es una forma viable de hacer aumento de datos en reconocimiento de emociones por voz. Incluso combinando bases de datos de diferentes lenguas, importante matiz, dado que la mayoría de los disponibles son en lengua inglesa.

También hemos observado que el clasificador estadístico usado (SVM) no logra grandes mejoras al aumentar el tamaño de las bases de datos. Por lo tanto, una vez comprobado que esta podría ser una línea viable para llevar a cabo el aumento de datos en reconocimiento de emociones, y puesto que con este experimento se ha establecido una línea base, la intención es probar este mismo método con los clasificadores y extractores de características que son el estado del arte en el procesado de voz. Para esto nos proponemos usar extractores de características que aprendan directamente de los datos, por ejemplo, Wav2Vec o Hubert, y clasificadores basados en redes neuronales, que pueden hacer un mejor uso del aumento de datos. En este caso, sería también interesante añadir más datasets al entrenamiento, ya que las redes neuronales tienen una capacidad mucho mayor que las SVMs de mejorar sus resultados al incluir más datos en el entrenamiento y más métodos de aumento de datos.

REFERENCIAS

[1]. Aaron Keesing, Yun Sing Ko, et al. “Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech”, Interspeech 2021.

[2]. B. Schuller, B. Vlasenko, et al. “Acoustic emotion recognition: A benchmark comparison of performances,” IEEE Workshop on Automatic Speech, 2009, pp. 552-557.

[3]. Mehmet B. Akçay and Kaya Oğuz. “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers”, Speech Communication, 2020, pp. 56-76.

[4]. F. Burkhardt, A. Paeschke, et al. “A database of German Emotional Speech”. Interspeech 2005. Technical University of Berlin.

[5]. Livingston, S.” The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)”. Ryerson University.

[6]. “The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database”. University of Southern California.

[7]. Florian, E. “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing”, Transactions on Affective Computing, vol 7 (2), 2015.

[8]. Felix Weninger, Florian Eyben, et al. “On the acoustics of emotion in audio: what speech, music, and sound have in common”. Journal Frontiers in Psychology, 2013.