UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

Departamento de sistemas informáticos y computación

Speech Emotion Recognition

Trabajo Reconocimiento automático del habla

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Autor
Francisco Javier Gil-Terrón Rodríguez

2021 - 2022

Tabla de contenidos

1. Introducción

En los últimos años, las técnicas de *sentiment analysis* y las áreas de estudio próximas, es decir, procesos por los que se identifica y extrae información subjetiva como sentimientos o emociones de los usuarios, han pasado de limitarse al ámbito académico a ser imprescindibles en el campo de *language understanding* cuyas aplicaciones cada vez están más presentes en los mercados, por ejemplo, analizando el *feedback* de los clientes sobre productos o servicios.

No obstante, esto hace referencia únicamente al análisis de textos, si añadimos el habla a la ecuación, se incrementa la complejidad. A esta tarea se le conoce como *speech emotion recognition o speech sentiment analysis*, en otras palabras, aplicar *sentiment analysis* directamente sobre la voz del usuario, lo que será crucial para mejorar los asistentes y sistemas de diálogo por voz actuales y progresar hacia las conversaciones 'naturales' artificiales. Es decir, el reconocimiento continuo de emociones mejoraría la interacción humano-máquina, lo que es un reto debido a la diversidad y complejidad tanto de las percepciones como de las mismas emociones.

Hoy en día esta es una tarea sobre la que se está trabajando y que no funciona de manera robusta fuera de dominios concretos, en parte por que los conjuntos de datos etiquetados con emociones no son demasiado abundantes y porque la mayoría de las aproximaciones son basadas únicamente en texto y no utilizan el reconocimiento automático del habla para mejorar en este sentido.

En este trabajo se hará una pequeña revisión sobre los principales problemas que se están abordando en la actualidad en materia de speech emotion recognition y cómo el estado del arte actual trata de resolverlos desde distintas aproximaciones.

Finalmente, se profundizará en una de estas aproximaciones para estudiarla con mayor detenimiento.

2. Problemas y propuestas actuales

2.1 Complejidad

El primero de estos problemas a los que se enfrenta el avance de *speech emotion recognition* es la complejidad que tienen de por sí los sistemas actuales de reconocimiento automático del habla, requiriendo de un profundo conocimiento en el área para poder ajustar correctamente los hiperparámetros de los modelos con el fin de que estos funcionen de manera adecuada.

Es por esta complejidad que, generalmente, la eficiencia de los modelos de lenguaje depende de la de los modelos acústicos debido a los errores en el proceso de reconocimiento automático del habla, provocando un efecto de cuello de botella.

Para tratar de solucionar esta problemática, y abordar este proceso de *fine-tuning* que puede llegar a ser muy costoso, un autor propone mejorar el *speech emotion recognition* tratando de minimizar el impacto del error del reconocimiento automático del habla mediante un mecanismo de *self-attention* combinado a una medida de confianza a nivel de palabra para reducir la importancia de las palabras con mayor probabilidad de haber sido reconocidas erróneamente y proporcionando en conjunto un mejor enfoque sobre las palabras que determinan el estado de ánimo [6].

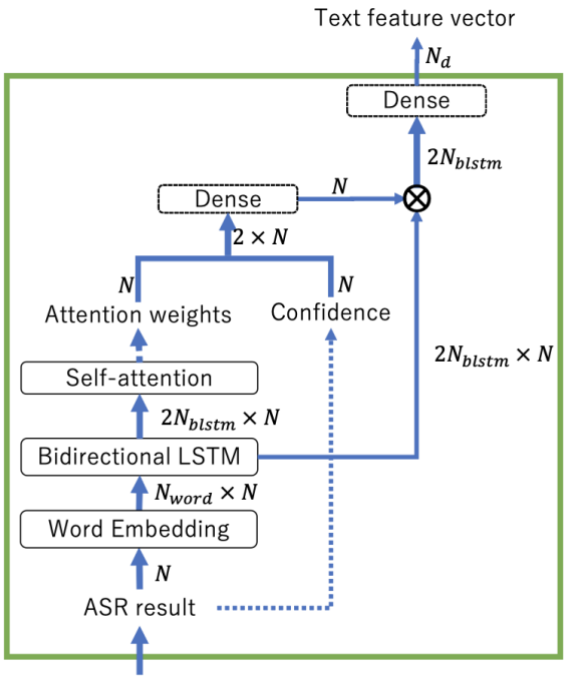

Figura 1 – Mecanismo de auto-atención con medida de confianza a nivel de palabra

En la figura anterior se puede ver la configuración que sigue dicho sistema, donde concatena esta medida de confianza con los pesos del mecanismo de self-attention para reducir la dependencia de las características textuales.

Por otro lado, otros autores proponen tratar configurar adecuadamente la parametrización con la precisión suficiente mediante la automatización de la optimización de los parámetros a través de otros sistemas y algoritmos (como *random forest* o redes bayesianas). En este caso, este proceso se aplica con el uso de una red neuronal convolucional con LSTM bidireccional (CNN-biLSTM) con lo que se alcanza un *accuracy* de 84% en la clasificación de sentimientos, eso sí, empleando muestras de audio obtenidas de películas, por lo que podría no ser robusto en un contexto real [3]. En la figura dos se ilustra la arquitectura de la red a la que se ha hecho referencia.

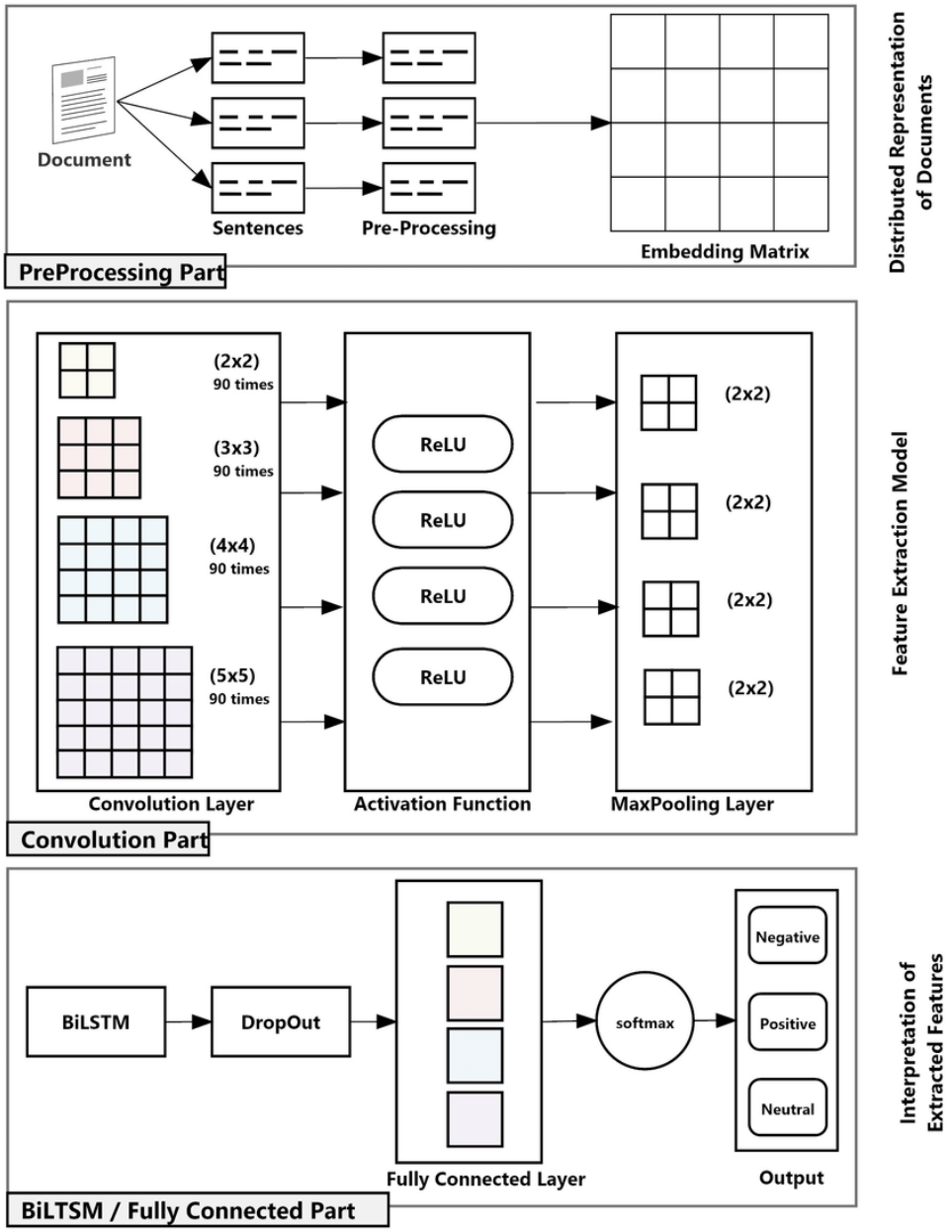

Figura 2 – CNN-biLSTM

2.2 Omisión de características acústicas

El segundo de los problemas que se han detectado es que, puesto que tradicionalmente el reconocimiento automático del habla se ha basado en la obtención de texto a partir de la voz, se omiten muchos de los elementos inherentes al habla humana que hacen que tenga una gran variabilidad, como el tono o el énfasis, con lo que se podrían mejorar los modelos acústicos actuales tal como sugieren algunos investigadores.

Es por esto por lo que los sistemas actuales son buenos en el contexto de reconocer habla 'limpia' pero tienen malos resultados en cuanto en cuanto se expande a habla espontánea.

Algunas propuestas para solventar este problema giran en torno a la utilización de recursos más allá de la propia de voz para identificar las emociones del usuario, y es cierto que disponer de imágenes sería de gran ayuda para determinar el estado de ánimo de un individuo, pero en la práctica, en la mayoría de las situaciones sólo se dispondría de la voz.

Por el otro lado, hay *papers* que proponen emplear modelos acústicos alternativos, como es el caso del que se ilustra en la siguiente imagen, con modelos acústicos basados en HMMs de trifonemas y mixturas de gaussianas o redes profundas, con lo que obtiene mejores resultados que los modelos del estado del arte, eso sí, en un dominio cerrado [5].

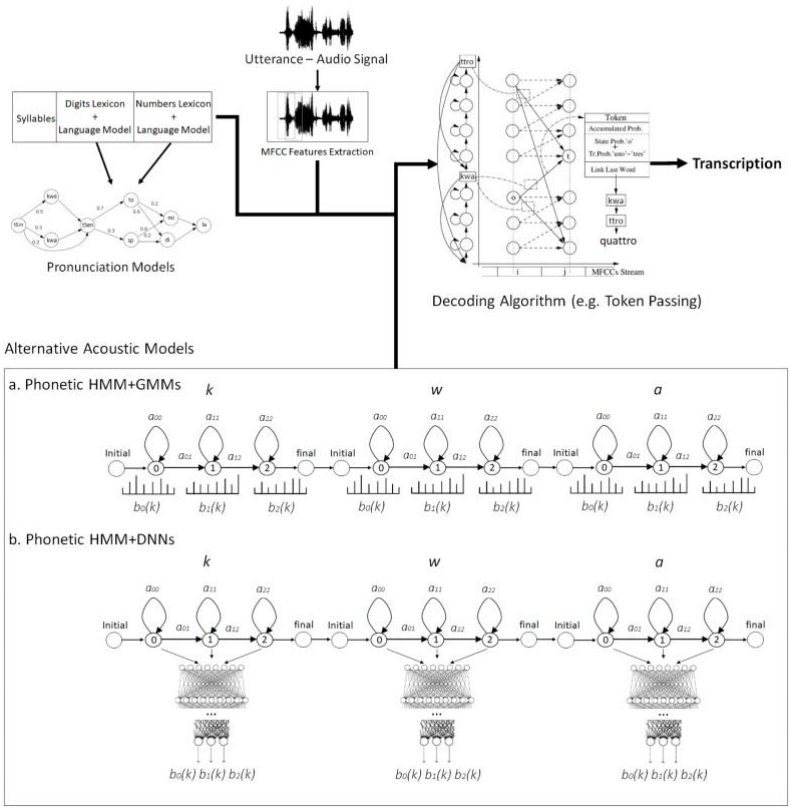

Figura 3 – Diagrama estándar de RAH con modelos acústicos alternativos

2.3 Escasez de datos etiquetados

Como en muchos otros campos, el uso *deep learning* se ha extendido y ha entrado ofreciendo alternativas y modelos nuevos que dan mejores resultados. No obstante, estos modelos necesitan una gran cantidad de datos para que funcione adecuadamente.

Siguiendo en esta línea, el tercer y último problema del que se hablará en este trabajo es la reducida cantidad de *datasets* que hay en la actualidad de información, tanto en forma de texto como de audio, etiquetada con emociones o sentimientos.

Además, existe una dificultad adicional, y es que generalmente las personas suelen mostrar sus sentimientos u opiniones (sobre todo de manera oral) en su lenguaje materno. Es por esto por lo que es más difícil obtener datos que pudieran resultar útiles de manera generalizada en países donde la lengua oficial no sea una mayoritaria como el inglés, chino o español.

Con el fin de contrarrestar esta problemática, artículos recientes plantean desde el uso de modelos de atención hasta técnicas de *transfer learning* para mejorar la clasificación de sentimientos o emociones con cantidades reducidas de muestras de entrenamiento [1],

Por otro lado, otros casos proponen abordar la mejora un modelo de *speech emotion recognition* empleando aprendizaje semi-supervisado en datos no etiquetados, de manera que se reevalúen iterativamente los datos etiquetados de manera automática para mejorar la confianza [2].

Finalmente, en el siguiente capítulo se tratará con más profundidad una alternativa diferente a las anteriores para este mismo problema, la escasez de datos etiquetados.

3. Alternativa más interesante

Como se ha venido diciendo, a continuación, se tratará con mayor detalle un caso en particular, el cual ha sido elegido porque, en mayor o menor medida, aborda los tres problemas anteriormente expuestos [7].

Este artículo parte de la premisa de que, aunque el lenguaje escrito y oral tengan características lingüísticas distintas, se pueden complementar para mejorar el entendimiento de las emociones. De esta manera, trata de crear un sistema capaz de extraer información sobre los sentimientos a partir de las señales acústicas a la vez que aprende sobre ello en la representación del texto.

Para solventar la hiperparametrización o *fine-tuning*, propone usar modelos pre-entrenados, que en este caso se hará uso de BERT, en un sistema convencional de *speech emotion recognition*, esto es, un proceso en dos pasos donde en primer lugar se realizará reconocimiento automático del habla y posteriormente se aplicará *sentiment analysis* con los modelos pre-entrenados.

Además de este, se propone un segundo planteamiento, complementario al primero, que tratará de enfrentar los otros dos problemas: la pérdida de características acústicas y prosódicas del reconocimiento automático del habla y la falta de información etiquetada en este ámbito. Para ello se utilizará una arquitectura end-to-end basada en aprendizaje semi-supervisado. Ambos planteamientos se pueden ver reflejados en la siguiente imagen:

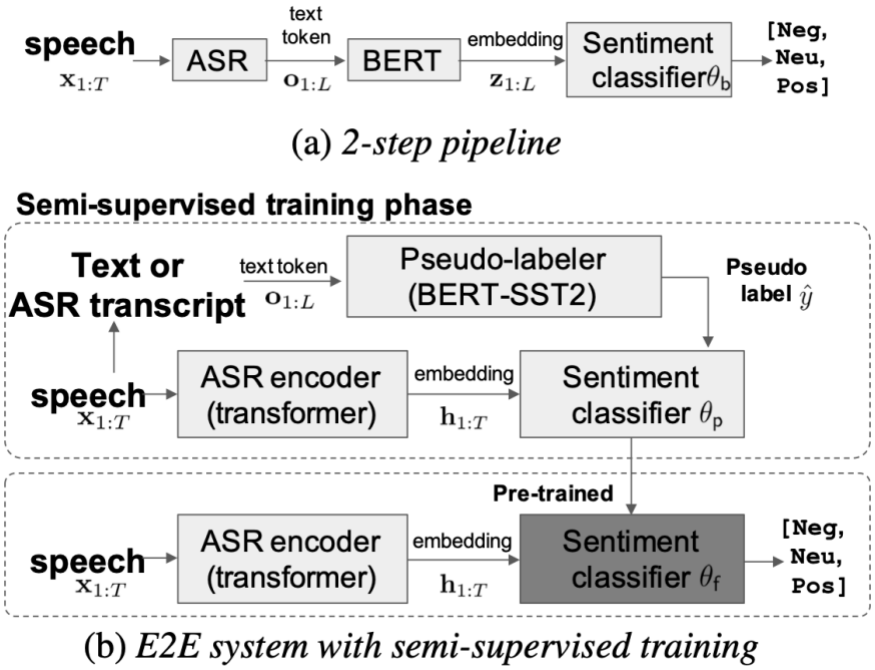

Figura 4 – Speech emotion recognition propuesto

En el caso de la propuesta end-to-end, se utiliza BERT como pseudo-etiquetador para los datos no etiquetados para entrenar un modelo de clasificación de sentimientos semi-supervisado con el que afinar el modelo de clasificación final que funcionará con las muestras etiquetadas.

Adicionalmente, en el artículo se afirma que con este planteamiento se alcanzó una reducción de hasta el 65% de supervisión humana aprovechando muestras no etiquetadas.

A continuación, en las siguientes tablas se muestran los resultados publicados por el autor contrastando las dos aproximaciones junto a valores de referencia (los embeddings son entrenados desde cero en lugar de emplear BERT), donde se puede ver que el sistema end-to-end obtiene los mejores resultados, además de obtener unos resultados bastante competitivos limitando el tiempo de entrenamiento a 5 horas con respecto a las 86 horas que consume el modelo final.

Architecture	SWBD-train transcript type	Validation Set (SWBD-test-ASR)						Evaluation Set (SWBD-holdout-ASR)					
		Unweighted			Weighted			Unweighted			Weighted		
		REC	PRE	F1	REC	PRE	F1	REC	PRE	F1	REC	PRE	F1
Baseline	GT	59.27	55.55	55.06	56.39	62.88	57.50	59.21	55.77	55.55	56.93	62.10	57.91
	ASR	52.57	50.07	47.38	47.60	58.44	48.30	52.43	49.84	47.66	47.92	57.12	48.48
	ASR (5h)	33.33	18.31	23.64	54.93	30.18	38.96	33.33	17.82	23.22	53.46	28.58	37.24
BERT	GT	**63.87**	64.64	**64.12**	**68.16**	68.01	**67.96**	**64.53**	65.05	**64.56**	**67.87**	**67.98**	**67.73**
	ASR	63.75	**65.21**	63.87	68.08	**68.46**	67.78	63.63	**65.13**	63.64	67.29	67.94	66.99
	ASR (5h)	50.18	55.01	50.99	61.08	58.88	58.82	50.09	56.06	51.03	60.85	58.91	58.34

Tabla 1 – Resultados del modelo tradicional en dos pasos

Fine-tuning dataset	Pseudo labeler	Semi-supervised training dataset	Validation Set (SWBD-test)						Evaluation Set (SWBD-holdout)					
			Unweighted			Weighted			Unweighted			Weighted		
			REC	PRE	F1	REC	PRE	F1	REC	PRE	F1	REC	PRE	F1
SWBD-train (86h)	-	-	64.59	**68.89**	**66.24**	**71.41**	**70.86**	**70.72**	61.21	65.92	62.74	67.73	67.89	66.99
	BERT-SST2	$\mathcal{S}$	63.68	67.65	65.23	70.37	69.79	69.71	62.37	66.68	63.85	68.47	68.58	67.85
	BERT-SST2	$\mathcal{S}, \mathcal{F}$	64.87	68.05	66.15	70.82	70.28	70.31	63.23	66.82	64.55	**69.05**	**68.77**	**68.46**
	XLNet-SST2	$\mathcal{S}, \mathcal{F}$	63.64	67.56	65.17	70.45	69.86	69.78	61.61	65.48	62.94	67.73	67.58	67.06
	BERT-SST2	$\mathcal{S}_{asr}, \mathcal{F}_{asr}$	**65.74**	66.51	66.11	70.23	70.01	70.10	**64.18**	65.28	**64.57**	68.27	68.35	68.14
SWBD-train (5h)	-	-	51.33	53.82	51.98	60.66	58.73	59.24	47.76	49.86	48.16	56.62	54.88	55.12
	BERT-SST2	$\mathcal{S}$	54.16	58.08	54.96	62.74	61.40	61.33	52.12	56.61	53.06	60.40	59.11	58.84
	BERT-SST2	$\mathcal{S}, \mathcal{F}$	**58.72**	58.67	**58.54**	**63.92**	**63.74**	**63.72**	**57.45**	**57.92**	**57.63**	**61.98**	**61.67**	**61.79**
	XLNet-SST2	$\mathcal{S}, \mathcal{F}$	58.19	57.89	58.00	62.63	63.07	62.82	56.86	57.39	56.75	60.59	61.52	60.74
	BERT-SST2	$\mathcal{S}_{asr}, \mathcal{F}_{asr}$	54.78	55.51	55.02	61.10	60.38	60.67	52.23	53.16	52.60	57.39	57.00	57.10

Tabla 2 – Resultados del modelo end-to-end

4. Referencias

[1] Sadr, H., & Nazari Soleimandarabi, M. (2022). ACNN-TL: attention-based convolutional neural network coupling with transfer learning and contextualized word representation for enhancing the performance of sentiment classification. *Journal of Supercomputing*. https://doi.org/10.1007/s11227-021-04208-2

[2] Li, Y. (2021). Semi-Supervised Learning for Multimodal Speech and Emotion Recognition. ICMI 2021 - Proceedings of the 2021 International Conference on Multimodal Interaction, 817–821. https://doi.org/10.1145/3462244.3481274

[3] Gumelar, A. B., Yuniarno, E. M., Adi, D. P., Sooai, A. G., Sugiarto, I., & Purnomo, M. H. (2021). BiLSTM-CNN Hyperparameter Optimization for Speech Emotion and Stress Recognition. International Electronics Symposium 2021: Wireless Technologies and Intelligent Systems for Better Human Lives, IES 2021 - Proceedings, 156–161. https://doi.org/10.1109/IES53407.2021.9594024

[4] Peng, Z., Dang, J., Unoki, M., & Akagi, M. (2021). Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech. Neural Networks, 140, 261–273. https://doi.org/10.1016/j.neunet.2021.03.027

[5] Coro, G., Massoli, F. V., Origlia, A., & Cutugno, F. (2021). Psycho-acoustics inspired automatic speech recognition. Computers and Electrical Engineering, 93. https://doi.org/10.1016/j.compeleceng.2021.107238

[6] Santoso, J., Yamada, T., Makino, S., Ishizuka, K., & Hiramura, T. (2021). Speech emotion recognition based on attention weight correction using word-level confidence measure. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 1, 301–305. https://doi.org/10.21437/Interspeech.2021-411

[7] Shon, S., Brusco, P., Pan, J., Han, K. J., & Watanabe, S. (2021). Leveraging Pre-trained Language Model for Speech Sentiment Analysis. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 1, 566–570. https://doi.org/10.21437/Interspeech.2021-1723

[8] Pappagari, R., Cho, J., Joshi, S., Moro-Velazquez, L., Zelasko, P., Villalba, J., & Dehak, N. (2021). Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 6, 4206–4210. https://doi.org/10.21437/Interspeech.2021-1850