POnSS, una nueva herramienta de segmentación discursiva

Rodd, J., Decuyper, C., Bosker, H. R., & Ten Bosch, L. (2021): A tool for efficient and accurate segmentation of speech data: announcing POnSS. *Behavior research methods*, 53 (2), pp. 744–756. doi:10.3758/s13428-020-01449-6

Reseña de Andrea Delgado Ortiz
Universidad de Alcalá

POnSS es una herramienta que procura simplificar el proceso de anotación o segmentación de un documento de audio. Hasta el momento, el software más utilizado es Praat y su herramienta integrada TextGrid. Sin embargo, la segmentación en Praat conlleva gran inversión de recursos humanos y temporales que no permiten seguir el ritmo vertiginoso de las humanidades digitales. Por ello, POnSS se presenta como la solución para obtener de manera automática resultados similares a los actuales.

La segmentación (ing. *speech segmentation*, o *speech alignment*) pretende separar de la cadena hablada elementos de menor tamaño, ya sean unidades léxicas, fonéticas o prosódicas. Realizada la segmentación, puede llevarse a cabo la anotación, la transcripción ortográfica o fonética que representa lo marcado en la cadena hablada. Los datos extraídos pueden constituir el corpus para un estudio posterior o la base de datos a partir de la cual se entrene un software de ASR[1] (ing. *Automatic Speech Segmentation*).

La segmentación en Praat se lleva a cabo con TextGrid, una herramienta que permite marcar en el espectrograma dónde está el límite de lo que se quiere separar y añadir texto plano de manera manual en la barra que aparece bajo los gráficos, que constituye la anotación. La segmentación y anotación en Praat requiere, pues, de la escucha repetida del archivo de audio para segmentar correctamente, y de una gran inversión de tiempo por parte del investigador previa al análisis de los datos.

[1] Los softwares de aplicación lingüística incluyen por defecto una herramienta para entrenamiento en la que, a partir de una base de datos previamente analizada por el lingüista computacional, la Inteligencia Artificial es capaz de reconocer patrones (sintácticos, fonéticos, del nivel que proceda según el caso y el software) para poder, posteriormente, buscar y reconocer por sí mismos patrones similares en entradas lingüísticas diferentes. Se trata, pues, de enseñar al programa.

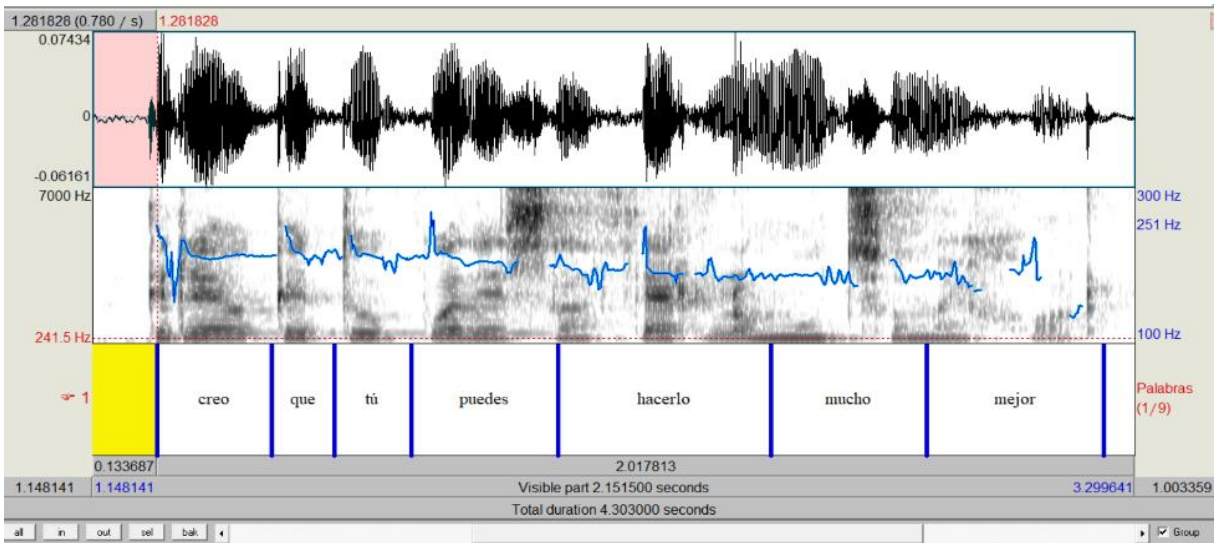

Figura 1. Ejemplo de espectrograma segmentado con anotación TextGrid. Fuente: elaboración propia.

Estos inconvenientes de la segmentación actual motivaron al equipo desarrollador de POnSS para crear una nueva herramienta que realice la misma tarea de manera automática y autónoma. Así, los autores consideraron tres dificultades o inconvenientes a superar.

En primer lugar, establecer una base sobre la que trabajar de manera cómoda. La segmentación manual se facilita en gran medida si la cadena ya está segmentada en palabras, por lo que no cierra oportunidades a análisis que tomen como unidad otros elementos distintos de la palabra.

En segundo lugar, mejorar la calidad de la segmentación en sí. Hasta el momento, los autores consideran que existe una correlación entre el tiempo invertido en la segmentación y el etiquetado y la calidad del estudio resultante (Rodd *et al.* 2021: 744), de modo que el ASR no es, en muchas ocasiones, una opción viable dado que la automatización reduce el tiempo empleado, pero también la calidad de los datos obtenidos.

En tercer y último lugar, facilitar la tarea de validación de la segmentación llevada a cabo de manera automática de modo que, aunque continúe siendo necesaria la mano de un investigador, su labor como auditor sea rápida y sencilla.

Para conseguir los objetivos, el software trabaja en tres fases: transcripción ortográfica, auditoría y revisión.

La transcripción ortográfica puede llevarse a cabo de forma manual y automática. Sin embargo, la automática no supone un ASR real, sino un *forced alignment*, es decir, en lugar de contar con un reconocimiento del habla se buscan las pausas y los *harmonicity peak*, que se corresponden con una vocal, que se vinculan con las palabras que se dan al programa, por lo

que asimismo es necesario contar previamente con una transcripción exacta del texto. En cuanto a la transcripción manual, la herramienta cuenta con la autonomía suficiente para segmentar en palabras, no así para etiquetar[2].

La parte de auditoría, *triage* para los autores, presenta las palabras sucesivamente en la pantalla: transcripción, oscilograma y espectrograma. La tarea del investigador es, simplemente, confirmar si la segmentación y la anotación son correctas. Podría considerarse una especie de entrenamiento NLU si la herramienta contara con Inteligencia Artificial. La sencillez de la interfaz de auditoría permite evaluar con rapidez un gran número de unidades segmentadas.

En la última parte, *retrimming*, se presenta de nuevo la palabra junto a su oscilograma y espectrograma. Sin embargo, en esta ocasión las palabras son aquellas que han sido rechazadas en el proceso de auditoría y se devuelven con un contexto mayor tanto en el audio como en los gráficos. Existen tres opciones para trabajar con estas muestras: aceptarlas, lo que supondría que la herramienta ha sido capaz de solventar por sí misma el problema; rechazarlas para que, de nuevo, la herramienta la devuelva con un margen mayor; o marcarlas para indicar que se precisa de intervención manual, por ejemplo, debido a un error de pronunciación.

Figura 2. Imágenes de las interfaces Transcription, Triage y Retrimming respectivamente. Tomado de Rodd *et al.* 2021: 747.

POnSS está abierto para descarga en GitHub. Sin embargo, al tratarse de una aplicación para Django[3], se dificulta la instalación y uso para investigadores que no precisan de conocimientos de programación, a diferencia de Praat. Su popularidad puede estar limitada, pues, por el *framework*[4] en que se ha desarrollado la herramienta.

[2] Se emplea el término *etiquetar* como sinónimo de *anotar*.
[3] Django: aplicación tipo *framework* que emplea Python como lenguaje de programación.
[4] *Framework*: entorno de trabajo que contiene previamente un lenguaje de programación y ciertas herramientas para facilitar el trabajo del desarrollador informático.

En cuanto al uso en sí de la herramienta, según indican los autores en la presentación, sí consigue simplificar el proceso de segmentación, pero siempre y cuando cumpla con el requisito de tener de antemano la transcripción, por lo que sigue siendo necesaria la escucha y transcripción manual con el gran costo temporal y humano que supone. Según los autores, el tiempo dedicado a la segmentación por parte del investigador se reduce en un 23% (Rodd *et al*. 2021: 753).

La principal ventaja que podría suponer, y que sin duda elevaría la popularidad – y el coste – de la herramienta sería dotarla de un motor de aprendizaje automático, de modo que las correcciones que realiza el investigador trasciendan el archivo con que se ha trabajado para darle al programa cierta retroalimentación para que mejore por sí mismo. Esto constituiría un proceso de auditoría real, y no una revisión como se realiza en la segunda parte de la herramienta.

Como conclusión, POnSS es una herramienta con potencial para convertirse en esencial para llevar a cabo tareas de segmentación, anotación y alineación si consigue ser realmente autónomo con un algoritmo de aprendizaje automático. Con ello, y el desarrollo como software o herramienta de Praat en lugar de su integración en un *framework*, se podrían vencer las limitaciones que tiene en la actualidad –menor flexibilidad y precisión– sin perder sus ventajas – mayor rapidez y eficiencia.

Referencias bibliográficas

Rodd, J., C. Decuyper, H. R. Bosker & L. Ten Bosch (2021): A tool for efficient and accurate segmentation of speech data: announcing POnSS. *Behavior research methods*, 53 (2), pp. 744–756. Recuperado de: https://doi.org/10.3758/s13428-020-01449-6.

Pleva, M., J. Juhár & A. Thiessen (2015): Automatic Acoustic Speech segmentation in Praat using cloud based ASR. *25th International Conference Radioelektronika (RADIOELECTRONIKA)*, pp. 172- 175. Recuperado de: https://doi:10.1109/RADIOELEK.2015.7129000.

Samant, R. M., M. R. Bachute, S. Gite & K. Kotecha (2022): Framework for Deep Learning-Based Language Models Using Multi-Task Learning in Natural Language Understanding: A Systematic Literature Review and Future Directions, in *IEEE Access*, vol. 10, pp. 17078-17097. Recuperado de: https://doi: 10.1109/ACCESS.2022.3149798.