UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

DESCRIPCIÓN TEMPORAL DEL DISCURSO E IDENTIFICACIÓN DE MOMENTOS COLABORATIVOS: DOS APLICACIONES DE LA INTELIGENCIA ARTIFICIAL AL ESTUDIO DE LA RESOLUCIÓN COLABORATIVA DE PROBLEMAS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

DANNER JORGE SCHLOTTERBECK MÉNDEZ

PROFESOR GUÍA:
PATRICIO FELMER AICHELE

MIEMBROS DE LA COMISIÓN:
ABELINO JIMÉNEZ GAJARDO
PABLO DARTNELL ROY
ROBERTO ARAYA SCHULZ

Este trabajo ha sido parcialmente financiado por el proyecto BASAL FB210005, CMM y la Escuela de Ingeniería y Ciencias, FCFM, Universidad de Chile.

SANTIAGO DE CHILE
2023

DESCRIPCIÓN TEMPORAL DEL DISCURSO E IDENTIFICACIÓN DE MOMENTOS COLABORATIVOS: DOS APLICACIONES DE LA INTELIGENCIA ARTIFICIAL AL ESTUDIO DE LA RESOLUCIÓN COLABORATIVA DE PROBLEMAS

La Resolución Colaborativa de Problemas (RCP) es una de las habilidades más importantes para el siglo 21 dada la complejidad de los problemas del mundo moderno. Sin embargo, a nivel internacional la mayoría de los estudiantes presenta un déficit de proficiencia en esta habilidad. Esto ha motivado un nuevo foco de investigación en RCP, generando nuevos instrumentos para analizar los aspectos cognitivos y sociales asociados a esta, los cuales se presentan usualmente en forma de rúbrica. En general, estos son usados por codificadores expertos para anotar las interacciones durante una sesión de RCP, comúnmente utilizando una grabación de vídeo o audio de la misma.

Este proceso, aunque extenso y detallado, toma demasiado tiempo, lo que imposibilita el analizar grandes volúmenes de sesiones. Producto de esto, investigadores han aplicado Inteligencia Artificial para facilitar el análisis de las sesiones. Sin embargo, trabajos previos utilizan equipamiento dedicado para la investigación, desde micrófonos y cámaras de alta definición, hasta plataformas computacionales especialmente diseñadas para mediar la RCP y sensores electrodermales. Esto merma la escalabilidad y flexibilidad de estos sistemas, dificultando el aplicar estas metodologías de forma masiva.

Con esto en mente, en este trabajo se desarrollan dos sistemas basados en Inteligencia Artificial para obtener representaciones temporales de sesiones de RCP, desde grabaciones de Zoom de las mismas. La primera representación temporal, obtenida desde el discurso de los estudiantes, identifica distintos tópicos en el discurso de las sesiones y los representa mediante sus palabras clave, para posteriormente describir la sesión a lo largo del tiempo en virtud estos. Por su parte, la segunda representación temporal, basada en la toma de turnos de los estudiantes durante la sesión, toma en cuenta que la colaboración no se presenta uniformemente a lo largo de la sesión y entrega dos índices automatizados de colaboración a lo largo del tiempo, permitiendo identificar los momentos de colaboración efectiva rápidamente.

Este trabajo presenta nuevas aplicaciones del aprendizaje no-supervisado al estudio de la RCP en forma de dos representaciones automatizadas para facilitar el análisis de las sesiones de aprendizaje colaborativo, sin necesidad de un equipo de codificadores y con requerimientos básicos de audio. Adicionalmente, las representaciones desarrolladas presentan un mecanismo para analizar de forma más general grandes cantidades de sesiones en poco tiempo. Por consiguiente, se espera que en un futuro estas puedan convertirse en una herramienta útil y valiosa para investigadores de esta área.

Agradecimientos

En primer lugar, quiero agradecer a los profesores Eugenio Chandia y Cristián Reyes por amablemente concederme acceso a los datos de su proyecto. Sin ellos esta tesis no hubiera sido posible. Quiero también agradecer al profesor Felmer por creer en este proyecto, y por siempre estar presente durante todo el proceso, incluso para prestarme un hombro en circunstancias difíciles.

En segundo lugar, quiero agradecer a los demás miembros de la comisión. A Abelino y Roberto, quienes me tomaron bajo su ala cuando yo aún no sabía ni conocía nada, me enseñaron, y me acompañaron a lo largo de todo este viaje. Además, también quiero agradecer a Pablo, por mostrar una voluntad de oro y un entusiasmo genuino con el proyecto desde el primer momento.

Por último, quiero agradecer a todos los amigos que formé durante este camino, tanto del DIM como de la Sección 5. Ustedes hicieron todo este proceso llevadero y memorable, y jamás me dejaron desistir las incontables veces que me lo cuestioné. A todos ustedes, les estoy por siempre agradecido y los llevo en el corazón.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Introducción

La Resolución Colaborativa de Problemas (RCP) se define como la capacidad de un individuo para involucrarse en un proceso en el cual dos o más agentes intentan resolver un problema compartiendo la comprensión y el esfuerzo, y aunando sus conocimientos, destrezas y esfuerzos para este fin [1]. Esta habilidad es considerada como una de las más importantes para el siglo 21, dado que la complejidad de los problemas en el mundo moderno conlleva la necesidad de articular equipos con múltiples áreas de conocimiento [2]. Adicionalmente, la RCP ha sido identificada como crítica para la eficiencia y la innovación en la economía global por la Organización para la Cooperación y el Desarrollo Económicos (OCDE) [1].

A pesar de la importancia de la RCP anteriormente mencionada, en la actualidad los niveles de competencia de los estudiantes en esta habilidad, a nivel internacional, se encuentran por debajo de lo esperado [3]. Esto ha motivado en la última década una nueva ola de investigación con respecto a los procesos cognitivos y sociales relacionados a esta habilidad a lo largo de distintos niveles educacionales. Investigadores han desarrollado diversas rúbricas que categorizan, cuantifican y califican las interacciones sociales dadas en una instancia de RCP, cada una teniendo múltiples sub-habilidades necesarias y formas de identificarlas. Ejemplos de esto pueden ser las desarrolladas por la OCDE, ATC21S [1], [4]. Sin embargo, existe un acuerdo general entre los investigadores en agrupar estas sub-habilidades en dos grupos.

El primero comprende las relacionadas al proceso social y colaborativo, como el establecer y mantener el conocimiento compartido del equipo, o el identificar las habilidades de cada integrante y asignar roles y tareas en función de esto. Por otro lado, el segundo grupo comprende las sub-habilidades relacionadas con la resolución de problemas, como el comprender, formular y explorar el problema, y la evaluación de la solución obtenida [5].

Así, las rúbricas descritas anteriormente permiten caracterizar los distintos tipos de interacciones necesarias entre los estudiantes para llegar a una solución del problema. Esto permite posteriormente estudiar diferencias en los tipos de interacciones entre distintos grupos en función de variables como el género, el estatus socio-económico, y la diversidad cultural y/o étnica [3]. Adicionalmente, también es posible estudiar las relaciones entre las interacciones y otras variables de interés como las ganancias de aprendizaje o los puntajes de pruebas estandarizadas. Mediante estos estos análisis se ha podido encontrar relaciones significativas entre las distintas habilidades relacionadas a la RCP y otras variables de interés, y desarrollar nuevas formas de enseñar para promover esta habilidad en los estudiantes [6].

En general, estas rúbricas son utilizadas por codificadores entrenados para anotar las interacciones presentes durante las sesiones de RCP, usualmente grabadas en video o audio. Por ende, si bien entregan información precisa y detallada con respecto a esta habilidad, este

proceso también tiene dos limitaciones. En primer lugar, el tiempo que toma anotar las sesiones es muy largo, puesto que los codificadores deben observar detalladamente cada sesión por separado, posiblemente repitiendo segmentos, y pausando para anotar cada interacción. Esto implica que el tiempo que se tarda en anotar una sesión puede llegar a ser varias veces la duración de la misma. En segundo lugar, dado que los codificadores deben estar entrenados en la rúbrica a ser usada, la fuerza de trabajo para esta tarea es reducida y requiere capacitación para ser expandida. Esto dificulta escalar estas metodologías y limita la capacidad de analizar nuevas sesiones en el futuro.

Debido a esto, se han realizado diversos intentos por facilitar el análisis de las interacciones de los estudiantes utilizando aprendizaje automático. Por ejemplo, en [7] se estudió la similaridad de las acciones realizadas por los estudiantes en un ambiente de RCP mediada por tecnología, encontrando que los grupos con similaridades crecientes a lo largo del tiempo presentaban las mayores ganancias de aprendizaje. En [8] se clasificó el trabajo de los grupos utilizando un índice de colaboración y se calcularon diversas métricas de cohesion del discurso, encontrando que varias de estas correlacionaban significativamente con el índice clasificado y las ganancias de aprendizaje de los alumnos. Por último, en [9], [10] y [11] se utilizaron distintos modelos de aprendizaje supervisado para clasificar las interacciones de los alumnos durante sesiones de RCP en distintas sub-habilidades como el compartir información, establecer conocimiento común, o monitorear el progreso del grupo.

Estos estudios presentan enfoques novedosos y muestran como el aprendizaje automático puede ser utilizado para facilitar el análisis de las sesiones de RCP, sin embargo, sus metodologías aún presentan algunas limitaciones. Por un lado, muchas de estas, y particularmente las que se basan en clasificar con respecto a alguna rúbrica en específico, se basan en el paradigma de aprendizaje supervisado. Esto implica que requieren de grandes cantidades de datos anotados por codificadores entrenados para obtener buenos desempeños sobre la tarea a automatizar. Además, dada la naturaleza supervisada de los sistemas desarrollados, estos son específicos para la rúbrica utilizada, por lo que de cambiar esta última dicho sistema necesitaría ser nuevamente entrenado. Por otro lado, en la mayoría de estos estudios se realizan en ambientes controlados y se utilizan sensores costosos y ocasionalmente invasivos para registrar las interacciones de los estudiantes, desde micrófonos de alta fidelidad hasta sensores electrodermales, de movimiento o de seguimiento ocular.

Esto dificulta el replicar estas metodologías a mayor escala, particularmente en zonas de bajos recursos y/o rurales. Con esto en mente, en esta tesis se proponen dos representaciones temporales no-supervisadas para sesiones de RCP obtenidas desde grabaciones de audio de las mismas. Estas representaciones están basadas en el contenido lingüístico del discurso de los estudiantes y la toma de turnos a lo largo de la sesión. Así, las representaciones propuestas no requieren de anotaciones previas ni equipamiento especial. Más aún, tampoco pretenden clasificar las interacciones de los estudiantes bajo alguna rúbrica en particular, sino que buscan ser una herramienta que facilite y agilice la exploración inicial a los investigadores en el área de RCP, permitiendo obtener análisis preliminares para un gran número de sesiones de forma escalable y sin necesidad de datos anotados.

Contexto y Descripción del Conjunto de Datos

Durante el desarrollo de esta tesis se aprovechó un conjunto de datos previamente levantado en el contexto del proyecto FONDEF "Modelo de desarrollo de habilidades matemáticas y de gestión de aprendizajes matemáticos para la Formación Inicial Docente en Educación Básica". El objetivo de este proyecto es activar las habilidades de resolución de problemas, argumentación, razonamiento, y aprendizaje colaborativo en los estudiantes de educación básica. Para esto, se busca promover en los futuros profesores el desarrollo de estrategias de enseñanza y aprendizaje centradas en los estudiantes, convirtiéndolos así en el principal motor de cambio para que esto ocurra.

Esto es llevado a cabo mediante la metodología de trabajo Colaborativo para la Enseñanza y Aprendizaje de la Matemática (CEAMA) en la formación inicial de los futuros docentes. La metodología CEAMA consta de elementos de selección y diseño de tareas para la formación inicial docente, y de gestión de las mismas en el aula. Las tareas utilizadas en esta metodología buscan generar conocimiento amplio y profundo respecto de la matemática escolar y su didáctica, es decir, permitir a los futuros docentes relacionar y conectar múltiples conceptos con distintos grados de abstracción, y plantear situaciones aplicables al aula que eliciten estas conexiones en los estudiantes.

Por su parte, el modelo de gestión del aula propuesto en CEAMA consta de cuatro fases. Primero, se separa a los estudiantes en grupos aleatorios y se les entrega el problema para que lo lean. Esto promueve la eliminación de barreras sociales dentro del aula y aumenta la movilidad de conocimiento entre los estudiantes al no estar presente el profesor. Segundo, el docente monitorea los grupos haciendo intervenciones cortas en forma de preguntas relacionadas al texto del problema en caso de ser solicitado o si lo estima necesario, sin esperar una respuesta por parte de los estudiantes. Esto se hace con el objetivo de fomentar la discusión entre el grupo y de guíar a los estudiantes en caso de ser necesario. Tercero, el profesor se asegura de que todos los estudiantes del grupo hayan comprendido la solución mediante preguntas a los distintos integrantes, y entrega una extensión del problema en dicho caso. Cuarto, se da una instancia de discusión planeada al final de la clase en la que los estudiantes del curso presentan sus soluciones y reflexionan sobre los conocimientos abordados.

A pesar de que la metodología CEAMA fue originalmente diseñada para implementarse en aulas presenciales, dada la situación sanitaria vivida durante los años 2020 y 2021 en Chile generada por el virus Covid-19, el proyecto fue llevado a cabo en modalidad remota. Esto presentó una oportunidad inesperada para, con el consentimiento y la ayuda de los estudiantes, grabar el trabajo de los grupos; permitiendo así realizar análisis más detallados a posteriori sobre las dinámicas presentadas durante las sesiones. Las clases fueron realizadas de forma semanal utilizando la plataforma Zoom, la cual permite al anfitrión de una reunión (profesor/a) agrupar aleatoriamente a los demás invitados en grupos de un tamaño predefinido y generar reuniones aisladas para cada grupo.

De esta forma, se logró levantar un total de 97 grabaciones (40 horas de audio) en las cuales los estudiantes trabajaron en problemas diseñados para activar las habilidades RCP. La metodología fue aplicada a tres cursos de la carrera de pedagogía en matemáticas en la Universidad de Concepción: Aprendizaje y Enseñanza de la Estadística y las Probabilidades

(Estadística), Aprendizaje y Enseñanza de la Geometría (Geometría), y Teoría Intuitiva de Números (Teoría de Números). La cantidad de grabaciones para cada curso fue de 38, 32, y 27 respectivamente; mientras que el largo promedio de las sesiones fue de 25 minutos (SD = 9 min). Además, el número de estudiantes en cada grupo podía variar en función del problema a abordar, con un mínimo de cuatro y un máximo de seis estudiantes.

Por último, se solicitó a los investigadores que originalmente habían levantado los datos el anotar un subconjunto de 9 sesiones, de forma de poder comparar los índices obtenidos con la anotación manual tras finalizar el desarrollo. La codificación manual de las sesiones se realizó utilizando una rúbrica desarrollada por los investigadores para medir distintas dimensiones del aprendizaje en grupo: *colaborativa, cognitiva* y *metacognitiva*. La dimensión colaborativa estudia los distintos tipos de interacciones que se dan entre los estudiantes en un nivel comunicacional, independiente de si estas se alinean con los objetivos de aprendizaje de la sesión. Por otro lado, la dimensión cognitiva estudia cómo se presentan y manejan dentro del grupo las ideas relacionadas a la resolución del problema a abordar. Por último, el nivel metacognitivo evalúa la capacidad de de auto-monitoreo de los individuos con respecto a si mismos y al grupo.

Adicionalmente, cada una de estas categorías contiene a su vez múltiples códigos los cuales categorizan las acciones que se pueden dar dentro de esta. La tabla a continuación presenta los distintos códigos presentes en esta dimensión de la rúbrica. De esta forma, las anotaciones manuales consistieron en una lista de frases anotadas con los tiempos de inicio y final de la frase en la grabación, la transcripción de esta, el alumno que habló, a quién estaba dirigida la frase, y el código de la acción correspondiente.

Descripción de las acciones clasificadas en la dimensión cognitiva de la rúbrica utilizada para codificar las sesiones.

Acción	Código	Descripción
Mencionar una Nueva Idea	MNI	Mencionar una nueva idea NO introducida antes.
Desarrollar una Idea	DUI	Elaborar, explicar o clarificar una idea, término, relación o información ya mencionada.
Aceptar una Idea	AUI	Aceptar o validar una idea en toda escala.
Rechazar una Idea	RUI	Rechazar o invalidar una idea en toda escala.
Hacer pregunta abierta	HPA	Hacer una pregunta para iniciar/continuar el debate y subir el nivel de discusión.
Comentario simple	COS	Hacer una pregunta o comentario para cerciorarse o confirmar algo, verificar.
Otros	OTR	Cosas que no clasifiquen en lo anterior.

Los capítulos siguientes presentan dos metodologías propuestas para obtener representaciones automatizadas de una sesión de RCP que ayuden a facilitar el análisis a los investigadores de este área. El primer capítulo presenta una representación automatizada del discurso

obtenida desde el audio que distingue de qué están hablando los alumnos a cada momento. Por su parte, el segundo capítulo presenta dos índices temporales automatizados para identificar momentos de colaboración productiva durante la sesión.

En ambos casos se comienza presentando una breve introducción al problema, luego se entrega una descripción detallada de la metodología y las técnicas utilizadas, seguido de la presentación y el análisis de los resultados obtenidos, y finalizando con la discusión de los mismos. Por último, se presentan conclusiones con respecto a las implicancias y limitaciones de los resultados obtenidos, y se proponen futuras lineas de investigación.

Capítulo 1

Desarrollo de una Representación Temporal No-Supervisada del Discurso

La serie de interacciones generada durante una sesión de RCP es la principal fuente de información para analizar las habilidades y conocimientos de un grupo [12]. Dentro de las múltiples formas en que estas interacciones pueden presentarse, el discurso, y particularmente la información lingüsística presente en este, ha sido utilizado exitosamente para modelar distintos procesos relacionados a la RCP como el argumentar, presentar ideas, y el negociar [13], [14], [10]. Además, existe evidencia de que la información lingüística presente durante la sesión también puede ser un buen predictor para otras variables de interés como la ganancia de aprendizaje de los alumnos y el desempeño del grupo en la tarea asignada [15], [16], [17]. Más aún, estos enfoques han sido posteriormente utilizados como herramientas para dar observaciones a los estudiantes con respecto al proceso colaborativo, identificar roles socio-cognitivos, y analizar dinámicas intra- e inter-personales a lo largo de las sesiones [18], [19], [20].

Así, estos estudios muestran que el contenido lingüístico del discurso es una buena fuente de información para automatizar análisis relacionados a la RCP. A pesar de esto, la investigación previa respecto a la automatización de dichos análisis se ha centrado principalmente en el paradigma de aprendizaje supervisado. Es decir, usualmente se busca etiquetar o evaluar cada interacción con respecto a una rúbrica previamente acordada y anotada, utilizando dichas anotaciones para que el sistema aprenda a etiquetar de forma automatizada. Esto implica que los sistemas obtenidos mediante estas metodologías están intrínsecamente atados a la rúbrica utilizada para entrenarlos, lo que los vuelve poco flexibles.

Con esto en mente, el objetivo de este capítulo es desarrollar un sistema que permita analizar las interacciones durante una sesión de forma no-supervisada, es decir sin necesidad de sesiones anotadas, basándose en alguna noción de similitud entre estas en vez de en una rúbrica en particular. Para esto se desarrolló un sistema, compuesto por varios modelos de aprendizaje automático en conjunto, el cual identifica los distintos contenidos latentes, llamados tópicos, en las sesiones y asigna cada interacción a uno de estos. De esta forma, las sesiones pueden ser posteriormente representadas a lo largo del tiempo basado en el tópico del que se habla a cada momento.

Adicionalmente, aprovechando que el conjunto de datos a utilizar está conformado por tres cursos distintos, se estudió las diferencias comportamentales entre estos en términos los tópicos más prominentes en cada uno. En particular, se encontraron tópicos que correspondían con los contenidos dictados en los distintos cursos y se verificaron resultados esperables con respecto a la proporción en la que estos aparecían en cada curso. Además, estos análisis permitieron encontrar diferencias significativas en la forma en que los estudiantes interactuaron en cada curso.

A continuación, la sección 1.1 presenta los distintos pasos para obtener los tópicos y la representación de cada sesión. Luego, en la sección 1.2 se presentan y analizan los tópicos encontrados, y se estudian las diferencias tendenciales entre las representaciones de los distintos grupos. Finalmente, en la sección 1.3 se discuten las limitaciones de la metodología y las implicancias de los resultados obtenidos.

1.1. Metodología

Esta sección detalla el procedimiento llevado a cabo para obtener una representación temporal que describa distintos tópicos y cómo estos se presentan a lo largo de una sesión de RCP, desde una grabación de audio de la misma. Este procedimiento, a grandes rasgos, consta de cuatro pasos. Primero, el audio de las sesiones es extraído de los videos y segmentado en frases cortas utilizando detección automática de voz. Segundo, las frases se transcriben utilizando un modelo de ASR (***A****utomatic* ***S****peech* ***R****ecognition*). Tercero, a cada frase se le asigna un vector en el espacio tridimensional utilizando técnicas de NLP (***N****atural* ***L****anguage* ***P****rocessing*) y reducción de dimensionalidad. Por último, se aplica mezclas de Gaussianas a los vectores tridimensionales para encontrar conjuntos latentes de frases en base a su cercanía espacial, llamados *tópicos*. Esto permite posteriormente analizar las sesiones temporalmente observando la variación de los tópicos en las distintas frases. Las subsecciones siguientes detallan las distintas fases de este proceso, mientras que la figura 1.1 resume el procedimiento completo.

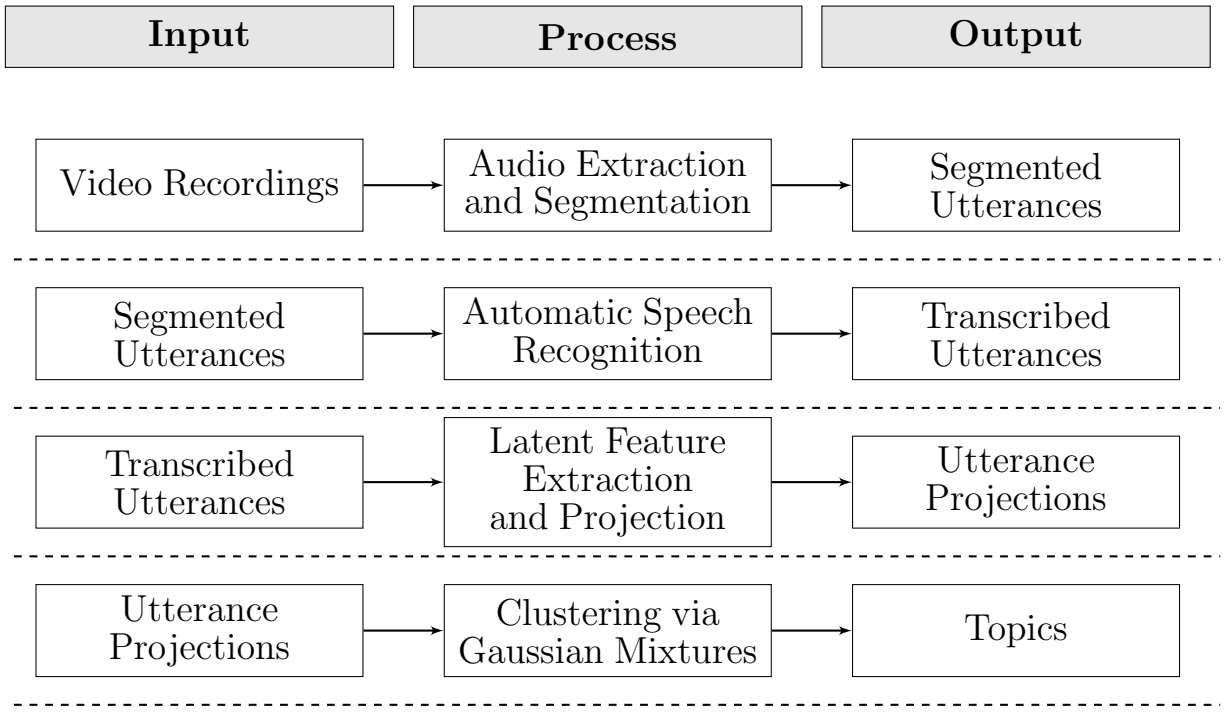


Figura 1.1: Pasos para la identificación de tópicos desde grabaciones en video de las sesiones.

1.1.1. Preprocesamiento de las Grabaciones

Antes de transcribir las grabaciones, se les aplicó una serie de transformaciones para prepararlas a un formato aceptable por el modelo de ASR. En primer lugar, se extrajo el audio de los videos de Zoom en formato *.wav*, se remuestreó a $16kHz$, y se convirtió a mono-canal. En segundo lugar, fue necesario dividir el audio en frases de pocos segundos para evitar problemas de memoria al transcribir y obtener un orden temporal que permita posteriormente analizar las sesiones a lo largo de esta dimensión.

Para esto, se utilizó un modelo de detección de voz o VAD (***V**oice **A**ctivity **D**etection*) con el objetivo de identificar los segmentos que contienen discurso. Estos modelos calculan propiedades acústicas de la señal sobre pequeñas ventanas móviles $(10 - 25ms)$ y luego asignan una salida binaria, como se muestra en la figura 1.2. De esta forma, se obtiene una secuencia de segmentos con voz $S = \{(t_k^i, t_k^f)\}_{k\in\{1,\dots,K\}}, K \in \mathbb{N}$ definida por los positivos consecutivos en la predicción, donde t_k^i, t_k^f marcan los tiempos de inicio y final del k-ésimo segmento. La segmentación fue llevada a cabo usando un modelo basado en redes recurrentes, desarrollado por el grupo de procesamiento de audio *pyannote* [21].

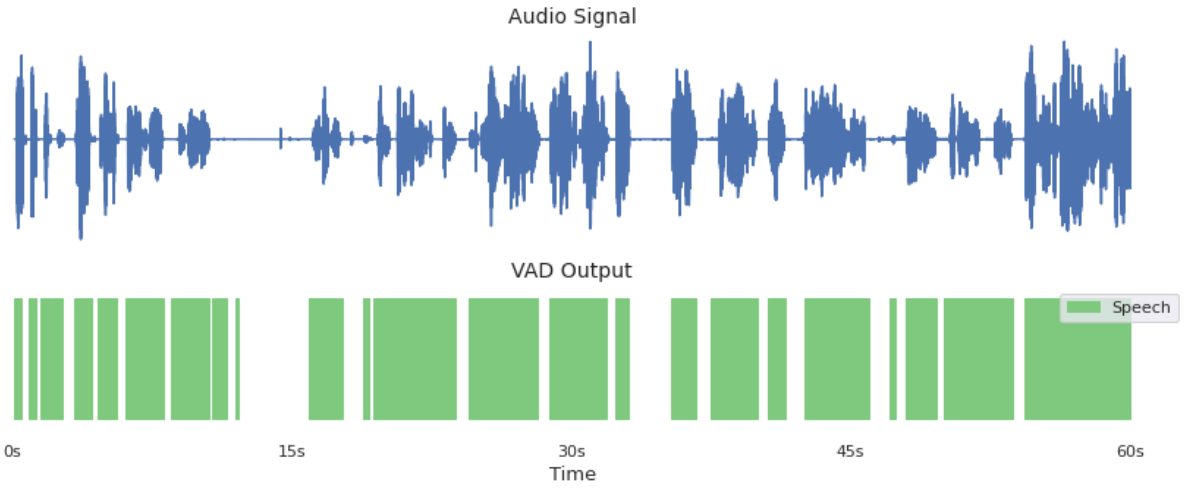


Figura 1.2: Ejemplo de VAD.

Posteriormente, se implementó un algoritmo sencillo de segmentación basado en los tiempos de inicio y final para extraer frases aproximando a un largo mínimo. Este algoritmo simula la forma en que los sistemas de ASR a tiempo real segmentan el audio (ver [22]), con la adición de que se busca generar segmentos de largo mayor a un parámetro t_{min}. Formalmente, dados t_{min}, el largo mínimo para considerar una unión de segmentos como una frase y w, el tiempo de espera para comenzar a contar una nueva frase; el algoritmo utilizado para obtener secuencia de frases segmentadas $\hat{S}$ desde S está dado por:

Código 1.1: Algoritmo de segmentación de frases basado en las fronteras dadas por un sistema de VAD.

input: S, t_{min}, w
output: $\hat{S} = \{(\hat{t}^i_j, \hat{t}^f_j)\}_{j=1}^m$
begin
  $\hat{S} \leftarrow \emptyset$
  $\hat{t}^i, \hat{t}^f \leftarrow t^i_1, t^f_1$
  for $k \in \{2, \ldots, n\}$:
    if $\hat{t}^f - \hat{t}^i > t_{min}$ **or** $t^i_k - \hat{t}^f > w$:
      $\hat{S} \leftarrow \hat{S} \cup \{(\hat{t}^i, \hat{t}^f)\}$
      $\hat{t}^i, \hat{t}^f \leftarrow t^i_k, t^f_k$
    else:
      $\hat{t}^f \leftarrow t^f_k$
    end
  end
  $\hat{S} \leftarrow \hat{S} \cup \{(\hat{t}^i, \hat{t}^f)\}$
  return $\hat{S}$
end

En este caso, el parámetro t_{min} regula el largo de los segmentos resultantes de forma de obtener frases con sentido suficiente para poder identificar tópicos. Por otro lado, el parámetro w regula que no se unan dos segmentos consecutivos demasiado separados, evitando mezclar dos frases con significados distintos y obtener segmentos demasiado largos. Esto permite evitar posibles problemas de memoria VRAM durante el entrenamiento del modelo de ASR debido al procesamiento de segmentos muy largos.

Sin embargo, una selección de w muy cercana a 0 resultaría en una segmentación en frases demasiado cortas para tener sentido o incluso palabras separadas en el caso de hablantes con baja cadencia. Por lo tanto, se probaron distintos pares de valores para t_{min} y w, buscando minimizar la cantidad de frases cortadas y/o erróneamente concatenadas. Con esto se estimó que valores apropiados para t_{min} y w se obtenían en 5 y 2 segundos, respectivamente, puesto que presentaron la mayor proporción de frases completas y la menor cantidad de segmentos erróneamente concatenados.

1.1.2. Transcripción del Audio Mediante Reconocimiento Automático del Discurso

Una vez segmentado el audio en frases, el siguiente paso para obtener una representación del discurso consta de transcribirlas. Para esto, se utilizó un modelo de ASR. En la actualidad, estos modelos suelen constar de dos partes. Primero, reciben una señal de audio y la segmentan en pequeñas ventanas, asignando a estas distribuciones de probabilidad sobre un alfabeto predefinido. Luego, las distribuciones obtenidas para cada ventana son decodificadas de forma de obtener palabras pertenencientes al lenguaje humano.

La principal ventaja de estos modelos frente a la transcripción manual recae en su escalabilidad, permitiendo procesar miles de grabaciones en poco tiempo y a bajo costo. A pesar de esto, dado que en este capítulo se busca desarollar un sistema que pueda ser posteriormente utilizado para analizar futuras sesiones de RCP, se optó por evitar el uso de transcriptores de la nube (e.g: Google, Azure, IBM), permitiendo evitar costos de escalabilidad y riesgos de privacidad sobre las grabaciones a analizar. Adicionalmente, los avances más recientes en el marco del ASR presentan resultados sobresalientes para esta tarea con mínimas cantidades de datos anotado, mientras estudios recientes sobre la aplicación de estos sistemas a audios de clases chilenas muestran un desempeño comparable al de sistemas disponibles en el mercado [23].

Con esto en mente, se decidió utilizar un modelo pre-entrenado de ASR que actualmente constituye el estado del arte, conocido como Wav2Vec 2.0. Este modelo logró su gran desempeño tras incorporar el paradigma de aprendizaje por transferencia a la tarea de transcripción automática, el cual consta a grosso modo de dos fases. En primer lugar el modelo aprende representaciones latentes y contextualizadas del audio mediante preguntas de tipo “llene el espacio en blanco” sobre grandes volúmenes de grabaciones no anotadas, proceso conocido como pre-entrenamiento o aprendizaje auto-supervisado. En la segunda fase, estas representaciones son utilizadas para entrenar la parte final del modelo, la cual asigna los caracteres a la señal de audio, sobre pocas grabaciones transcritas, proceso conocido como fine-tuning. A continuación se describen los criterios utilizados para seleccionar el modelo pre-entrenado, aplicarle fine-tuning y decodificar la salida de este.

1.1.2.1. Selección y Fine-tuning del Modelo de ASR

Una de las grandes ventajas del proceso de aprendizaje por transferencia, es que reduce considerablemente la cantidad de datos anotados necesarios para lograr una transcripción aceptable al aprovechar lo aprendido por el modelo en la fase de pre-entrenamiento. Adicio-

nalmente, esto también permite adaptarlo rápidamente a dialectos e incluso idiomas distintos sin necesidad de re-entrenar toda la arquitectura, via fine-tuning. Sin embargo, esto también implica que la elección del modelo pre-entrenado es particularmente importante si sólo se busca hacer fine-tuning, dado que las representaciones aprendidas son el punto de partida para la transcripción. En particular, en esta investigación se decidió utilizar una versión multilingual de Wav2Vec 2.0 pre-entrenado por Jonatas Grosman sobre el subconjunto de habla hispana del corpus *CommonVoice* [24], [25].

Esta versión fue seleccionada dado que presenta los mejores resultados sobre el conjunto de prueba de *CommonVoice* en la tabla de paperswithcode.com [26]. Adicionalmente, esta decisión fue motivada por el hecho de que *CommonVoice*, al ser un corpus de construcción colectiva, presenta un rango mucho más variado de calidades de audio, acentos y dialectos a diferencia de otros corpus como *LibriSpeech* o *VoxPopuli* [27], [28]. Esto lo vuelve ideal, considerando la variada calidad de audio, como también el discurso más relajado y espontáneo que se presentan en las sesiones de RCP.

Una vez definida la versión pre-entrenada del modelo, se seleccionó una muestra de 692 frases de las obtenidas durante el preprocesamiento (∼2 horas y 20 minutos de audio en total) para ser transcritas manualmente y ser usadas para el proceso de fine-tuning. Tras seleccionarlas, las muestras transcritas manualmente fueron divididas aleatoriamente en proporciones 60/20/20, generando tres conjuntos con 83, 29, y 28 minutos de audio, respectivamente. Estos conjuntos fueron usados para entrenar la parte final del modelo utilizando un enfoque de entrenamiento-validación-testeo [29].

Es decir, se aplicó fine-tuning al modelo pre-entrenado sobre el conjunto de entrenamiento (60 %), pero buscando minimizar la función de pérdida CTC (***C****onnectionist* ***T****emporal* ***C****lassification*) sobre el conjunto de validación (20 %) para evitar sobreajustar el modelo [30]. Este proceso fue repetido inicializando el modelo con múltiples combinaciones de hiperparámetros y finalmente la combinación que presentó mejor desempeño en términos de WER [31] sobre el conjunto de validación fue seleccionada. Por otra parte, el conjunto de testeo (20 %) fue aislado para posteriormente generar una comparación contra un transcriptor de la nube. La implementación fue llevada a cabo utilizando la librería `Transformers` [32].

1.1.2.2. Modelo de Lenguaje

Dado que los sistemas de ASR procesan pequeñas ventanas de audio ($\sim 20ms$) y asocian cada una a una letra, no tienen realmente noción de si el texto decodificado corresponde efectivamente a palabras existentes, ni de si la transcripción tiene sentido humano. Para sortear estas dificultades, los modelos de ASR usualmente son usados en combinación con modelos de lenguaje, los cuales buscan complementar la falta de entendimiento del lenguaje humano del transcriptor.

En este caso, se utilizó un modelo de bigramas entrenado sobre el texto del conjunto de entrenamiento. Estos modelos se basan en el conteo de la aparición de pares de palabras sobre un corpus vasto de texto, para luego asignar la siguiente palabra a decodificar basado en la probabilidad condicional sobre la palabra anterior. Los bigramas fueron entrenados utilizando la librería de C++ `KenLM` y acoplados al modelo de ASR entrenado utilizando la librería `pyctcdecode` para decodificar la salida de este último [33], [34]. Finalmente, el transcriptor

desarrollado, compuesto del modelo de ASR en conjunto con el de bigramas, fue usado para transcribir las frases de todas las sesiones disponibles.

1.1.3. Modelamiento del Contenido del Discurso Mediante Modelos de Tópicos

Tras transcribir las grabaciones de las sesiones, el siguiente paso apunta a contestar la pregunta *¿De qué se está hablando durante la sesión?*. Para ello, se utilizaron *modelos de tópicos*. Esta técnica modela una frase como una distribución de probabilidad sobre un conjunto de tópicos. A su vez, un tópico τ es una distribución de probabilidad sobre un vocabulario, donde la probabilidad de la palabra v en el tópico τ es considerada una medida de representatividad. De esta forma, es posible interpretar un tópico mediante sus palabras con mayor representatividad, y una frase como una mezcla de distintos tópicos.

El procedimiento para mapear las frases transcritas en distribuciones de probabilidad sobre un conjunto de tópicos consta de 4 fases: análisis exploratorio y preprocesamiento de los datos, embebimiento contextualizado del texto en un espacio de alta dimensión, proyección sobre un espacio de baja dimensión, y computación de las distribuciones de los tópicos. A continuación se describe cada una de estas fases en detalle.

1.1.3.1. Análisis Exploratorio y Limpieza de los Datos

En primer lugar, es necesario llevar a cabo una limpieza de los datos para asegurar el buen uso de las técnicas posteriores. Para esto, se comenzó por estudiar la distribución de la duración de los segmentos transcritos, encontrando una cantidad importante de segmentos de menos de $1s$, cuya frecuencia decaía exponencialmente hasta los $5s$, seguido de un nuevo máximo en el intervalo $[5, 6]s$ con frecuencias que disminuyen exponencialmente en los intervalos siguientes, como muestra la figura 1.3.

Dado que la cantidad de información posiblemente contenida en intervalos de tan corta duración es dudosa, se decidió estudiar una muestra aleatoria del intervalo $[0, 5]s$. Esto reveló que una gran parte de los segmentos en dicho intervalo correspondían a falsos positivos del sistema de detección de voz, causados principalmente por ruido ambiental o roces con el micrófono. Más aún, se observó que las transcripciones de dichos intervalos correspondían sistemáticamente a (a) repeticiones de un único caracter (e.g: "a a a a a") o (b) concatenación una o más vocales (e.g: "ooaaao").

Sin embargo, durante el muestreo, también fueron encontrados segmentos que contenían efectivamente discurso de los alumnos y cuyas transcripciones eran válidas y útiles para el análisis. En consecuencia, la posibilidad de una limpieza basada únicamente en la duración de los segmentos fue considerada contraproducente debido a la posible pérdida de datos útiles para el modelo. Adicionalmente, tras realizar un muestreo del intervalo $[5, \infty)s$, se pudo observar también casos de falsos positivos en la detección de voz en este intervalo, los cuales presentaban los mismos patrones en la transcripción.

Con lo anterior en mente, se optó por una heurística que buscara identificar dichos patrones, basando la limpieza en la transcripción misma en vez de la duración de los segmentos.

En particular, se utilizó un filtrado por diccionario, utilizando la lista de las 10.000 palabras más frecuentes publicada por la Real Academia Española (RAE). Además, se agregó al diccionario el chilenismo "po", dado que durante los análisis exploratorios se observó una gran frecuencia de este, y se removieron las palabras de una sola letra debido al primer comportamiento sistemático antes mencionado.

Posteriormente, se estudió los segmentos cuyas transcripciones tuvieran dos o menos palabras pertenencientes al diccionario anteriormente descrito. Esto reveló que dichos segmentos correspondían en su mayoría a los mismos patrones descritos en párrafos anteriores con la excepción de segmentos que correspondían a una palabra aislada del resto del discurso (por al menos dos segundos antes y después, dada la heurística de segmentación), los cuales fueron considerados prescindibles dada su baja frecuencia. Finalmente, se calculó la proporción de tiempo transcrito que fue removido del conjunto de datos, obteniendo una estimación de falsos positivos en la detección de voz del 16.4 %. La figura 1.3 muestra el histograma de los datos antes y después de la limpieza.

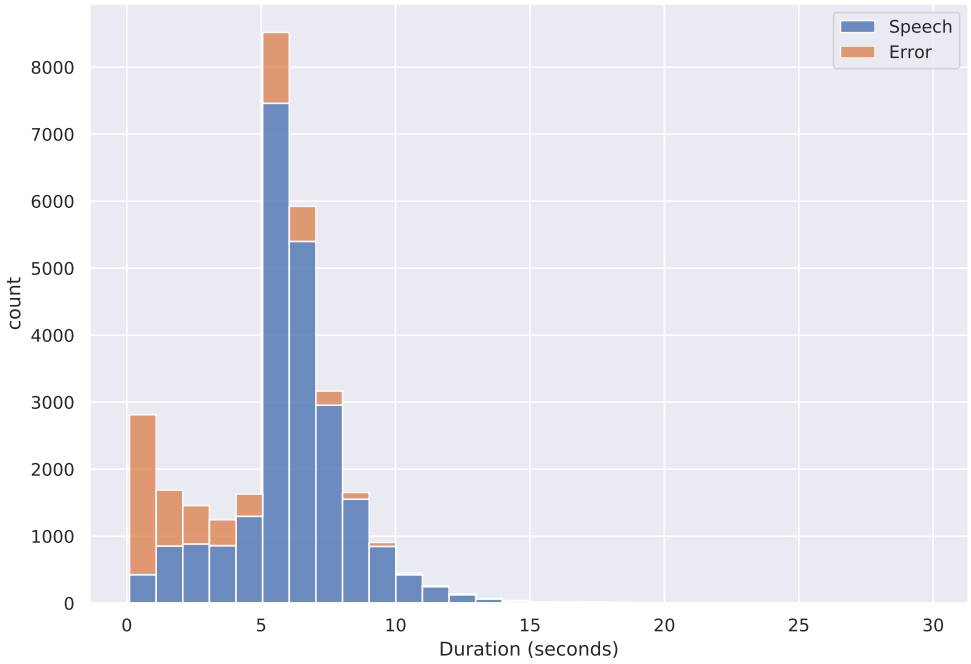


Figura 1.3: Histograma de la duración de las frases clasificadas como error (naranjo) y discurso (azul).

1.1.3.2. Mapeo de las Frases en Vectores Latentes Contextualizados

El primer paso para modelar el discurso es encontrar una representación numérica del texto que nos permita posteriormente aplicar técnicas de aprendizaje no supervisado. Para lograr este objetivo se utilizó el modelo RoBERTa, el cual constituye al momento de la escritura el estado del arte en NLP, alcanzando el mejor rendimiento en múltiples tareas del

punto de referencia GLUE [35].

Similar a Wav2Vec 2.0, RoBERTa está basado en el paradigma de aprendizaje por transferencia, es decir, aprende representaciones generales desde grandes cantidades de datos rellenando palabras en blanco, para luego usarlas en tareas más específicas como clasificación de texto o identificación de pregunta-respuesta. Dado que se busca una representación latente que capture las dependencias entre las palabras en el lenguaje, se decidió aprovechar la salida de las capas pre-entrenadas de RoBERTa, es decir, las representaciones generales sin aplicarles *fine-tuning* a alguna tarea en específico.

Para obtener el vector contextualizado de una frase transcrita, primero se debe transformar cada palabra de esta en una serie de *tokens* pertenecientes al vocabulario de RoBERTa utilizando el *tokenizador* del modelo. Luego, a cada token se le asocia un vector contextualizado de dimensión 768 usando las capas pre-entrenadas del modelo. Finalmente, los vectores de cada token que componen la frase son agregados para obtener un vector contextualizado de esta última.

De esta forma, las frases transcritas obtenidas con el modelo de ASR de la sección anterior fueron mapeadas en un espacio latente que captura los significados de las palabras que la componen y sus dependencias. Para la implementación, se utilizó una versión en español, RoBERTuito, entrenada por el grupo Chileno de análisis de sentimientos *pysentimiento* [36]. RoBERTuito fue entrenado usando más de 500 millones de tweets con un alto porcentaje del corpus de procedencia chilena. Debido a esto, se espera que pueda capturar de mejor manera el lenguaje presente durante las sesiones de RCP.

1.1.3.3. Reducción de Dimensionalidad

Una vez las frases han sido mapeadas en el espacio latente, el siguiente paso para encontrar los tópicos es agruparlas en dicho espacio, de forma que sea posible encontrar similitudes semánticas en las mismas. Sin embargo, dada la alta dimensión del espacio de salida de RoBERTa, es necesario primero proyectar los vectores en un espacio de menor dimensión para evitar la llamada *maldición de dimensionalidad* al calcular las distancias entre las frases [37], [38]. Esto fue hecho utilizando el algoritmo UMAP (***U**niform **M**anifold **A**pproximation and **P**rojection*) [39].

Este algoritmo aproxima los datos a una variedad Riemanniana de la dimensión deseada y luego mapea esta última a un espacio euclídeo. Este proceso se logra mediante la computación de un grafo particular de vecinos más cercanos, el cual es posteriormente dibujado en el espacio euclídeo para obtener las proyecciones de los datos. En general, los principales hiper-parámetros de este modelo corresponden a la dimensión de la variedad, la distancia mínima entre dos puntos en la proyección, la métrica de similaridad utilizada en el espacio de salida, y la cantidad de vecinos considerados para calcular el grafo, siendo esta última la que regula si se prioriza la estrucutra topológica local de los datos (pocos vecinos), o la global (muchos vecinos).

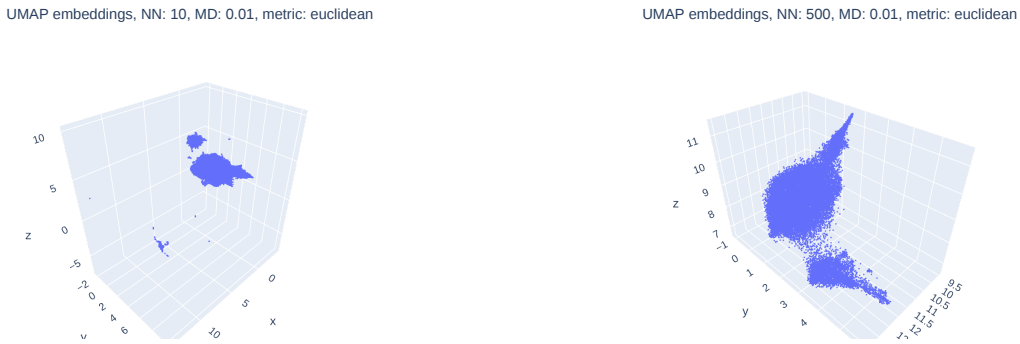
Durante esta investigación, las frases fueron proyectadas en una variedad tridimensional para facilitar la posterior visualización de los datos. Adicionalmente, dado que se busca clusterizar las proyecciones, se buscó utilizar distancias mínimas en la proyección cercanas a

0, pues valores mayores tienden a perder estructura local y equiespaciar las proyecciones, haciendo más difícil clusterizar los puntos. De forma similar, la distancia debería ser estrictamente mayor que 0 para evitar que los puntos muy cercanos en el grafo tengan la misma proyección. Con esto en mente, se fijó el valor de la distancia mínima en 0.01.

Por otro lado, para el número de vecinos y la métrica de similaridad a utilizar para calcular el grafo no se tenían razones para escojer valores particulares, por lo que se optó por la exploración y la visualización de las proyecciones para decidir los valores a utilizar. En el caso de la métrica, se consideraron dos que son comunes para los algoritmos de reducción de dimensionalidad: el inverso aditivo de la distancia euclidiana y la similaridad de coseno, mientras que para el número de vecinos se consideró la grilla $\{10, 25, 50, 100, 200, 500\}$.

Posteriormente, se generaron visualizaciones interactivas de las proyecciones con las distintas combinaciones de parámetros utilizando la librería `plotly` y se estudiaron cualitativamente sus propiedades. Durante el análisis se buscó evitar puntos aislados en la proyección y observar si este heredaba alguna separación visible desde el espacio inicial. Estas características son deseables pues evitan que posteriormente se presenten tópicos muy pequeños y/o difusos para interpretarlos. Las visualizaciones interactivas están disponibles para descarga en el siguiente enlace.

Durante el análisis de las proyecciones se observó que estas, a pesar de una reducción al espacio tridimensional, presentaban una separación visible en dos clusters, siendo uno notoriamente de mayor tamaño. Más aún, se observó que las transcripciones correspondientes al cluster más pequeño correspondían a frases cortas (a lo más 4 palabras), mientras que las encontradas en el cluster más grande tenían un largo de 5 o mayor. Adicionalmente, se observó que en los valores más bajos del número de vecinos la proyección era demasiado localizada, lo cual generaba puntos aislados (figura 1.4.a), mientras que en los valores más altos la estructura global de los puntos era muy preponderante, hasta el punto de llegar a unir los dos conjuntos encontrados (figura 1.4.b).

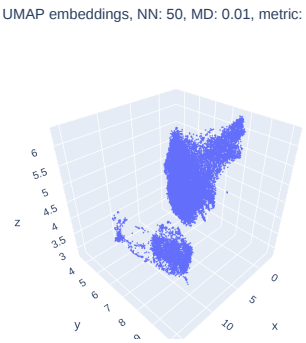


(a) Proyecciones obtenidas con $NN = 10$

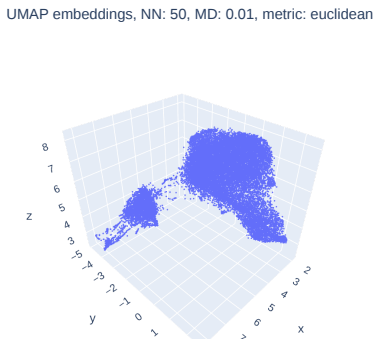
(b) Proyecciones obtenidas con $NN = 500$.

Figura 1.4: Proyecciones tridimensionales obtenidas utilizando UMAP con la métrica euclidiana para 10 (izquierda) y 500 (derecha) vecinos (NN). Se observan los outliers en el primer caso y la pérdida de estructura local en el segundo.

Por otro lado, ambas métricas presentaban comportamientos similares para los distintos valores del número de vecinos y una separación similar en dos clusters principales basados en el largo de la frase. Sin embargo, se observó una tendencia en el caso de la similaridad de coseno a posicionar los dos clusters anteriormente mencionados en planos cuasi-paralelos (figura 1.5.a). Finalmente, tomando en consideración lo descrito en los párrafos anteriores, se decidió utilizar 50 vecinos para generar las proyecciones y utilizar la métrica euclidiana, pues esta configuración permitía obtener un buen balance entre la estructura local y global (figura 1.5.b). Esto permitió posteriormente facilitar al algoritmo de mezclas de gaussianas a encontar clusters a lo largo de las tres dimensiones.



(a) Proyecciones obtenidas con $NN = 50$ y similaridad de coseno.



(b) Proyecciones obtenidas con $NN = 50$ y métrica euclidiana.

Figura 1.5: Proyecciones tridimensionales obtenidas utilizando UMAP con la similaridad de coseno (izquierda) y la métrica euclidiana (derecha) para 50 vecinos (NN). Se observa que se comienzan a proyectar los puntos de los dos subconjuntos en planos cuasi-paralelos.

1.1.4. Obtención de los Tópicos via Mezclas de Gaussianas

Una vez que se han obtenido las proyecciones tridimensionales de cada frase, estas fueron agrupadas en tópicos basado en su cercanía espacial. Para esto, se usó una mezcla de gaussianas sobre el espacio proyectado. Es decir, cada tópico $\tau \in \mathcal{T}$ es modelado como una normal multivariada $\mathcal{N}(\mu_\tau, \Sigma_\tau)$ con probabilidad de pertenencia π_τ. Así, la probabilidad de observar una frase dadas las distribuciones de los tópicos puede ser modelada como una combinación convexa de $\{\mathcal{N}(\mu_\tau, \Sigma_\tau)\}_{\tau \in \mathcal{T}}$:

$$\mathbb{P}(x|\pi, \Sigma, \mu) = \sum_{\tau \in \mathcal{T}} \pi_\tau \mathcal{N}(x|\mu_\tau, \Sigma_\tau), \quad \sum_{\tau \in \mathcal{T}} \pi_\tau = 1, \quad \pi_\tau > 0$$

donde la normal multivariada está dada por

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\},$$

De esta forma, se busca maximizar la verosimilitud del conjunto de frases proyectadas X con respecto a la familia $\{(\pi_\tau, \mu_\tau, \Sigma_\tau)\}_{\tau \in \mathcal{T}}$, dada por:

$$\prod_{x \in X} \mathbb{P}(x|\pi, \Sigma, \mu)$$

1.1.4.1. Selección del Número de Tópicos

El parámetro más importante al clusterizar utilizando mezclas de Gaussianas es, sin duda, el número de estas a modelar, el cual en nuestro caso corresponde al número de tópicos subyacentes en el conjunto de frases. Por esto, considerando que a priori no conocemos el número de tópicos subyacentes en el conjunto de frases, se decidió utilizar una heurística basada en el BIC (***B**ayesian **I**nformation **C**riterion*) [40]. Este criterio de selección mide la ganancia de información de un modelo y penaliza por la cantidad de parámetros del mismo. Así, un decrecimiento en el BIC implica que existe una mayor cantidad de información de los datos retenida en las distribuciones de las Gaussianas.

De esta forma, se estimó un número apropiado de tópicos calculando el BIC sobre la grilla $k \in \{2, \ldots, 100\}$ y luego estableciendo un umbral (en este caso del 5 %) para la diferencia relativa entre el BIC calculado para cada k y el mínimo acumulado hasta la iteración previa. De esta forma, se obtiene una heurística similar al conocido *método del codo* utilizado en el algoritmo *K-means* [41], a la vez que se limita la cantidad de tópicos con el umbral para facilitar el análisis. Tras aplicar la heurística descrita, se obtuvo un valor de 13 tópicos para este conjunto de datos. La figura 1.6 presenta el BIC calculado y el mínimo acumulado a lo largo de la grilla.

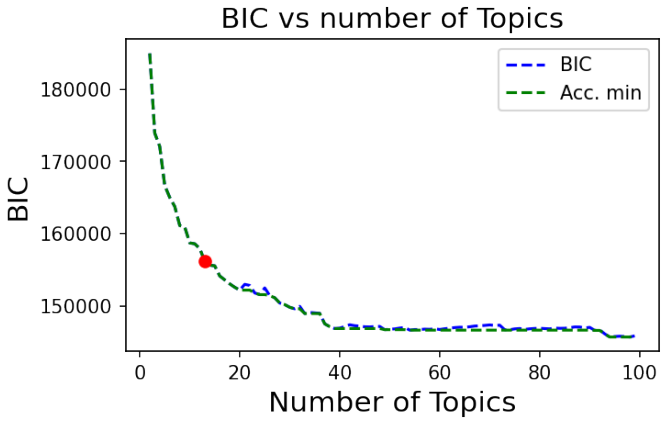


Figura 1.6: Curva de BIC y mínimo BIC acumulado utilizando GMMs con $k \in \{2, \ldots, 100\}$.

1.1.4.2. Interpretación de los Tópicos via TF-IDF

Luego, para obtener las distribuciones de probabilidad de cada tópico sobre el vocabulario, se utilizó (c-)TF-IDF (*[**c**lass-based] **T**erm **F**requency - **I**nverse **D**ocument **F**requency*) [42]. Este método entrega una noción de importancia de cada palabra para cada tópico, basado en la frecuencia de la palabra en dicho tópico y penalizando por la cantidad de otros tópicos en los que aparece dicha palabra. Formalmente, sean $\mathcal{T}$ los tópicos, para una frase $x \in X$, definimos por $\#(x)$ la cantidad de palabras (con reposición) de x. Análogamente, si V es

el vocabulario presente en la colección de tópicos, para una palabra $v \in V$ definimos $\#_v(x)$ como la cantidad de veces que v aparece en x. Entonces la importancia de la palabra v en el tópico $\tau \in \mathcal{T}$ está dada por:

$$tfidf(v, \tau, \mathcal{T}) = \underbrace{\frac{\sum_{x \in \tau} \#_v(x)}{\sum_{x \in \tau} \#(x)}}_{tf(v,\tau)} \cdot \underbrace{log\left(\frac{|X|}{\#_v(X)}\right)}_{idf(v,\mathcal{T})}$$

Donde $\#_v(X) = \sum_{x \in X} \#_v(x)$. Así, es posible caracterizar un tópico τ por sus palabras con mayor $tfidf(v, \cdot, \mathcal{T})$, lo cual nos entrega una noción interpretable de su contenido. De forma similar, tras dotar de interpretación a los tópicos, es posible analizar cada frase como una mezcla de estos, dado el modelo utilizado. Esto permite construir una representación de los tópicos a lo largo de la sesión basado en cómo varían las distribuciones de estas contra los tópicos a lo largo del tiempo.

1.2. Análisis de Resultados

En esta sección se presentan y describen los resultados obtenidos en los distintos pasos de la metología aplicada. Comenzando por los resultados obtenidos en la evaluación del sistema de ASR y algunos ejemplos de transcripciones, continuando con un análisis y visualización de los tópicos obtenidos, y finalizando con una representación de las sesiones en base a los tópicos y un estudio de diferencias entre los tópicos y los distintos grupos de sesiones.

1.2.1. Evaluación del Sistema de ASR

Para tener un punto de referencia al evaluar la calidad del modelo, se hizo uso de la plataforma Google Cloud (GCP), la cual es una de las plataformas más utilizadas de ASR a nivel mundial. Esta plataforma fue seleccionada dado que otorga créditos de prueba tras la suscripción. Aprovechando esto, se transcribieron los segmentos del conjunto de testeo utilizando el sistema *Speech-to-Text* de GCP.

Luego, el modelo que presentó el mejor desempeño en el conjunto de validación en términos de WER y CER (***W****ord* y ***C****haracter* ***E****rror* ***R****ate*, respectivamente) fue probado y comparado contra la transcripción del servicio *Speech-to-Text* de *Google* sobre el conjunto de testeo. Adicionalmente, se calcularon las mismas métricas para Wav2Vec2.0 sin modelo de lenguaje, para tener una segunda linea base. En la tabla 1.1 se aprecian los valores obtenidos sobre el conjunto de testeo por los distintos sistemas. Además, se calculó un T-test sobre la media de las diferencias de cada métrica en las distintas frases con respecto al modelo desarrollado, donde la hipótesis nula correspondía a una media igual a 0.

Tabla 1.1: Valores de WER y CER obtenidos por los distintos sistemas, y reducciones porcentuales con sus respectivas significancias estadísticas del T-test. (*: $p < 0.01$, **: $p < 0.001$)

Modelo	**WER**	**CER**	**ΔWER %**	**ΔCER %**
Wav2Vec2+LM	51.2 %	31.1 %	-	-
Wav2Vec2 base	60.4 %	30.2 %	15.2 %**	−3.0 %
Google S2T	54.6 %	39.4 %	6.2 %	21.1 %**

Como se puede observar en la tabla, el transcriptor entrenado es competitivo con los transcriptores encontrados en el mercado, presentando mejoras relativas de un 6.2 % con respecto a *Speech-to-Text* y un 15 % con respecto al modelo base sin decodificador. También se observa un pequeño degrade en CER con respecto al modelo base, probablemente producto de que el modelo de lenguaje a veces asocia palabras erróneas con tal de ajustarse al vocabulario. Sin embargo, dicho degrade no resulta ser estadísticamente significativo. Además, se logra una mejora importante en CER con respecto al sistema de *Google*, a la vez que se logran transcripciones con sentido las cuales pueden ser posteriormente procesadas con RoBERTa, como puede verse en la tabla 1.2

Tabla 1.2: Ejemplos de transcripciones obtenidas usando los distintos sistemas de ASR.

Transcripción Manual	Wav2Vec2 + LM	Google S2T	Wav2Vec2 Base
"sí yo creo que sí sí igual es importante todo todo lo que habíamos hecho"	"si yo creo que sí porque igual son es importa te todo todo lo que haya ocho"	"sillones y v igual son importante todo todo lo que hayamos hecho"	"siño creo que ssí porque igual son es importante todo todo lo que hayamo echo"
"se puede comer toda la circunferencia que da hasta veinticinco metros entonces yo lo hice por fuera de la cerca pero es por dentro de la cerca"	"se puede comer toda la circunferencia que da hasta dos como cinco meto por de yo lo hice por fuera de eso que por dentro de"	"se puede comer toda la circunferencia queda hasta dos como cinco metros por fuera de la sda"	"se puede comer toda la circunferencia que da hasta dos como cinco metros epor dez yo lo hice por fuera delesefrosa busqiapero por dentro de"
"será necesario copiarla ahí como información en la pizarra yo creo que sí"	"será necesario copiar la ahí como información en la pizarra yo creo que como era en eso"	"era necesario copiar la y como información en la pizarra"	"será necesario copiarla ahí como información en la pizarraeya creo que como soperaen eso"

1.2.2. Análisis de los Tópicos Obtenidos

Tras obtener los tópicos, se procedió a analizarlos de forma cualitativa para comprenderlos antes de representar las sesiones temporalmente. Esto se hizo mediante el análisis de las palabras características y las frases con mayor probabilidad de pertenencia a cada tópico. Adicionalmente, se utilizó una visualización interactiva de los tópicos en el espacio proyectado para estudiar donde estaban ubicados espacialmente. A continuación se detallan los análisis realizados sobre los tópicos.

1.2.2.1. Análisis de Palabras Características

En primer lugar, se interpretaron los tópicos mediante un estudio de las palabras más representativas de cada uno. Las palabras fueron seleccionadas basadas en su *tf-idf* con respecto a cada tópico, de esta forma obteniendo un ranking desde la más importante a la menos importante. A grandes rasgos, se encontraron dos tipos de tópicos: los que denotaban interacción entre los estudiantes y elaboración de ideas en conjunto, con palabras como "si", "entonces", "po", "creo", "pregunta", etc; y los que correspondían a un contenido en espécífico, compuestos de palabras del estilo "ángulo", "triángulo", "rectángulo", "polígono", etc.

Basados en esta dicotomía, se decidió etiquetar los topicos que denotaban interacción entre los alumnos con la etiqueta INTER. Adicionalmente, dentro de los tópicos de contenido se identificaron claramente dos temas. Los primeros correspondían a tópicos de geometría, tanto básica ("triángulo, "rectángulo, "cuadrado") como más avanzada ("paralelas, "recta , "punto"), a los cuales se etiquetó como GEOM. Por otro lado, el segundo grupo mostraba tópicos de números ("ciento", "treinta", "cincuenta") a los cuales se les denominó NUM. También se encontraron algunas palabras relacionadas a estadística como "datos" y "gráfico", sin embargo, no fueron aísladas mayoritariamente en un tópico a diferencia de las anteriores.

Por otro lado, dentro de los tópicos de INTERoración de ideas sobre el problema, fue más dificil encontrar diferencias significativas entre ellos debido a una alta superposición de palabras. No obstante, se encontraron algunos tópicos como el INTER6 y el INTER7, de los cuales es posible inferir información acerca de la acción que se estaba realizando (compartir pantalla y trabajar el problema en la pizarra, respectivamente). La tabla 1.3 muestra las palabras con mayor importancia para los tópicos descritos anteriormente.

Tabla 1.3: Palabras con mayor *tf-idf* para 6 de los tópicos obtenidos. Se observan los tópicos descritos en los párrafos anteriores

GEOM1	**GEOM2**	**NUM2**	**INTER2**	**INTER6**	**INTER7**
si	ángulo	ciento	profe	voy	pizarra
dos	ángulos	treinta	pregunta	ver	grupo
paralelas	triángulo	cincuenta	si	compartir	entonces
igual	dos	cuarenta	gráfico	pantalla	vamos
recta	lado	veinticinco	datos	poner	si
ejemplo	rectángulo	tres	hacer	dejar	ver
ser	pelo	veinticuatro	creo	si	problema
dice	si	cinco	entonces	va	hacer
puede	tres	seis	igual	puedo	ser
entonces	polígono	ochenta	conocimiento	hacer	podría
mismo	lados	sesenta	ejemplo	buscar	puede
punto	recto	cuatro	niños	intentar	dos
creo	corto	siete	podríamos	deja	así
podría	color	veintidós	problema	puede	podemos
rectas	arriba	ocho	preguntas	ahora	podríamos
puntos	cuadrado	dos	ser	volver	pregunta
así	po	veinte	dice	ir	poner
número	negro	veintiuno	clase	entonces	informe
plano	cuadrilátero	nueve	podría	aquí	creo
misma	abajo	dieciseis	puede	correo	después

1.2.2.2. Análisis de Frases Más Representativas de Cada Tópico

En segundo lugar, se estudió los tópicos a través de sus frases con mayor probabilidad de pertenecia obtenida desde la GMM. Para esto, se analizaron las 10 frases más representativas de cada tópico, para complementar los hallazgos encontrados con respecto al contenido de los tópicos utilizando *tf-idf*. En general, las frases más representativas presentaron una pertenencia fuerte respecto a sus tópicos asociados, encontrándose en todos los casos sobre el 96 % de probabilidad, y bajo el 98 % en sólo dos tópicos de INTER. Adicionalmente, para presentar los resultados de una manera más intuitiva, se generaró una visualización interactiva de las frases proyectadas, lo cual permitió estudiar los tópicos con respecto a su posición en el espacio.

Tras observar las frases con mayor probabilidad de pertenencia a cada tópico, el primer hallazgo fue que las frases en INTER1, INTER3, INTER5 y NUM3 eran notoriamente más cortas que en otros tópicos, siendo en general de 5 palabras o menos. De estos, se encontró los tópicos INTER1, INTER5 y NUM3 eran los que conformaban el cluster pequeño observado tras proyectar con UMAP, mientras que INTER3 era el más cercano espacialmente a los tres anteriores. Adicionalmente, se observó que el tópico INTER3 correspondía a frases que eran cortas y presentaban falsos positivos en el sistema de detección de voz, causando transcripciones similares a las observadas durante el análisis exploratorio (ver tabla 1.4). La

figura 1.7 presenta las proyecciones de la frase, mientras que la su versión interactiva está disponible para descarga en el siguiente enlace.

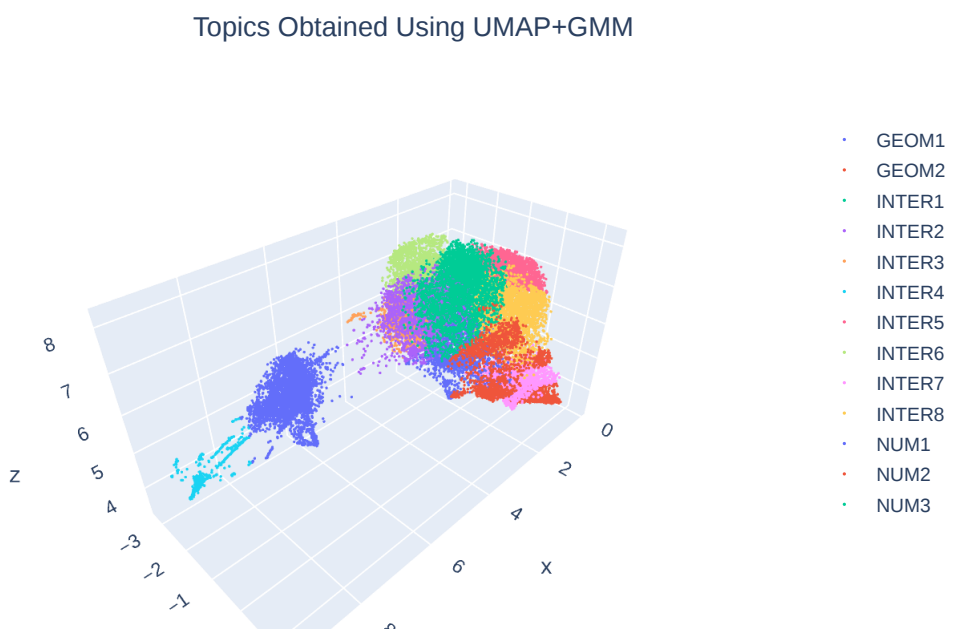


Figura 1.7: Proyecciones de las frases en el espacio tridimensional, etiquetadas con los distintos tópicos.

Esto reveló que los tópicos no sólo fueron separados por el contenido del texto de la frase, sino que también por la cantidad de palabras de la misma. Por añadidura, este resultado explica la existencia de tópicos con palabras características similares, dado que aunque sus *tf-idf* sean cercanos, las frases asociadas fueron separadas espacialmente durante la proyección debido a su largo, dando origen a dos tópicos distintos, pero con palabras similares (ver INTER4 e INTER5 en tablas 1.5 y 1.4). Sin embargo, y a diferencia de los tópicos pertenecientes al cluster más grande, las frases encontradas en los tópicos INTER1, INTER3 y INTER5 eran en general poco entendibles y difíciles de interpretar dado que eran muy cortas (tabla 1.4).

Tabla 1.4: Frases con mayor porcentaje de pertenencia a los tópicos INTER1, INTER3, INTER Y NUM3

INTER1	INTER3	INTER5	NUM3
"ya lo varios que"	"la de a a a"	"primero medio"	"cuatro siete cuatro siete seis"
"creo que en esa las"	"la razón a a a"	"gráfico circular"	"cuatro siete y cuatro ocho"
"yo creo que no y"	" a va a va a"	"seis siete"	"cinco seis cinco siete seis"
"ya super no a gracias"	"a las antes a a"	"escala le"	"un tres cuatro cinco seis"
"ya creo que lo gracias"	"es un en a a"	"poco tampoco"	"siete cinco cinco cinco"

Por el contrario, se encontró que las frases pertenecientes a los tópicos restantes se correspondían fuertemente con los contenidos asignados a cada tópico utilizando sólo sus palabras más representativas. Es decir, las frases asociadas a tópicos NUM y GEOM corroboraron los distintos contenidos inferidos desde las palabras claves, mientras que las frases pertenecientes a los tópicos INTER denotaban una clara interacción entre los estudiantes al mostrar frases dirigas claramente a un compañero. Más aún, algunos tópicos INTER denotaban formas particulares de interacción, como comunicar/corroborar una opinión con respecto a como abordar el problema o avisar que se va a hacer una acción por cuenta propia (ver INTER2 e INTER6 en tabla 1.5)

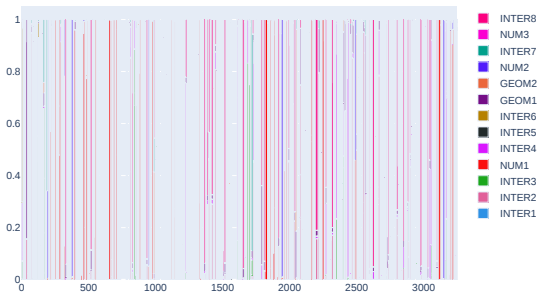
Tabla 1.5: Frases con mayor porcentaje de pertenencia a los tópicos INTER2, INTER4, INTER6 Y GEOM1

INTER2	INTER4	INTER6	GEOM1
"y un mensaje breve obviamente no tiene que ser un correo super extenso porque tampoco tenemos tiempo para estar escribiendo"	"de la esta hay también po porque la prueba le había dijo y como que despues cacho"	"ya lo voy a anotar en el chat para que"	"en plano estaban solamente las dos direcciones que sería largo y el ah pero en el espacio son tres el largo el no y el alto osea a lo que le decía de las dos dimensiones y el otro de las tres dimensiones"
"lo que podemos hacer que estamos ciento era darle hasta cada interpretación resultado y decir es la profesora"	"mejor esa porque de investigación y los niños de segundo la pueden respondievan a decir no se po porque a rico conce"	"ya voy a abrir de nuevo me lo quiero la cuenta porque"	"congruentes dos figuras geométricas son congruentes si tienen las mismas dimensiones y la misma forma sin importar su posición u orientación"
"porque no trata de hace lo que más podamos de por un video youtubatard lo hacemos por que ahora"	"o sea así como bien si po porque esta bien hecho ejercicio necesitan eso"	"voy voy a buscar a ver que un geoplano porque clara"	"y tampoco hay un solo plano porque ya no estan en el mismo plano no hay un punto de intersección y tampoco estan en un mismo plano"
"yo creo que estamos o no faltaría algo más yo creo que tenemos lo que"	"si si po yo creo que es es cierto a la por eso po"	"donde sale ya voy a es perame voy a sacar un pantalla esto y voy"	"pero en el espacio a parte de que no se corte tienen que existir en los mismo plano"
"pero esperemos que el porque le preguntamos que tenemos que hacer con eso que"	'y por al menos por lo que esos van y poner eso po al plano"	"voy a buscar los puntos a ver si tengo algo que ir yo estoy viendo al que sea"	"cambio explica que recta en el plano si dos rectas son paralelas no hay intersección y en la imagen a intersección toda decir que son paralelas"

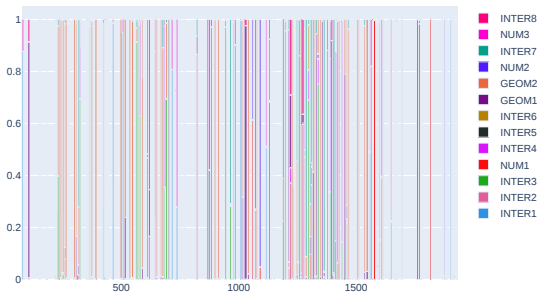
1.2.3. Representación de las Sesiones y Análisis Estadístico sobre los Grupos

Tras el análisis de los tópicos, se buscó representar la dinámica de los mismos a lo largo del tiempo durante las sesiones. Para esto, se generó una linea de tiempo donde cada frase estaba representada por una barra coloreada con la distribución de probabilidad de los tópicos sobre la frase. Como era esperable, dada la naturaleza de los grupos, se observó que las sesiones correspondientes a Teoría de Números y Geometría presentaban una mayor proporción de tiempo en estos tópicos. Ejemplos de esto se pueden encontrar en la figura 1.8 mientras que las versiones interactivas de las imágenes para cada sesión pueden ser encontradas en el

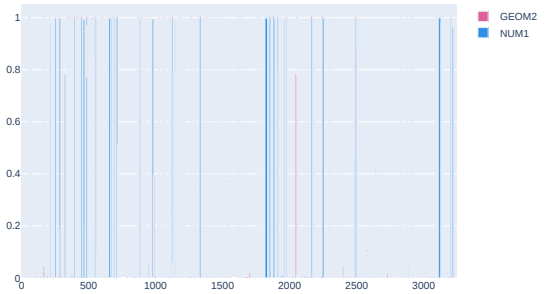
siguiente enlace.



(a) Ejemplo clase de Teoría de Números, todos los tópicos.

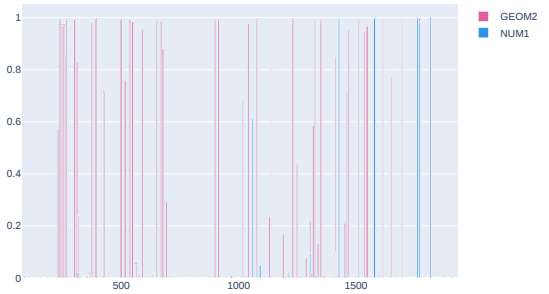


(b) Ejemplo clase de Geometría, todos los tópicos.



ing [MathJax]/extensions/MathMenu.js

(c) Ejemplo clase de Teoría de Números, sólo NUM1 y GEOM2.



(d) Ejemplo clase de Geometría, sólo NUM1 y GEOM2.

Figura 1.8: Ejemplos de las distribuciones de los tópicos a lo largo de clases de Geometría y Teoría de Números. Se observa la preponderancia de los tópicos GEOM2 y NUM1, respectivamente.

Dado que los patrones observados sugieren una diferencia en la proporción en que los tópicos se presentan para los distintos cursos, se decidió estudiar esta relación de forma más detallada. Para esto, se generaron tablas de contingencia representando la cantidad de frases pertenecientes y no pertenecientes a cada tópico, con respecto a cada curso. Luego se estudió la independencia entre la pertenencia al tópico y el curso mediante un test χ^2. La tabla 1.6 muestra la proporción de frases pertenecientes a cada tópico para los distintos cursos, y el $p-$valor obtenido en el respectivo test.

Tabla 1.6: Proporción de frases pertenencientes a cada tópico para los distintos cursos y $p-$valor resultante del test de independencia. (*<0.05, **<0.01, ***<0.001)

Tópico	**% Estadística**	**% Geometría**	**%T. de Números**	$p - val$
INTER1	10.2 %	12.6 %	8.0 %	***
INTER2	13.9 %	4.1 %	10.6 %	***
INTER3	15.6 %	17.4 %	13.9 %	***
INTER4	9.6 %	9.3 %	9.7 %	-
INTER5	2.5 %	3.2 %	2.1 %	***
INTER6	2.6 %	3.1 %	2.3 %	**
INTER7	10.6 %	7.7 %	9.1 %	***
INTER8	7.1 %	4.6 %	9.6 %	***
NUM1	6.0 %	5.9 %	**12.3 %**	***
NUM2	3.4 %	1.2 %	**4.9 %**	***
NUM3	4.3 %	5.3 %	3.4 %	***
GEOM1	9.3 %	**12.8 %**	10.3 %	***
GEOM2	4.7 %	**12.6 %**	3.6 %	***

Como se observa en la tabla, se encontraron diferencias significativas en las proporciones de todos los tópicos a lo largo de los distintos cursos, con excepción del tópico INTER4. Adicionalmente, se pudo comprobar lo estipulado en párrafos anteriores, encontrando una prominencia estadísticamente significativa de los tópicos GEOM y NUM en las sesiones de Geometría y Teoría de Números, respectivamente. Además, se observó que las clases de Geometría presentaban una mayor proporción en los tópicos INTER1, INTER3, INTER5 y NUM3; los cuales corresponden a los tópicos de frases cortas.

En contraste, se observó en promedio una menor proporción de los tópicos INTER2, INTER4, INTER7 y INTER8 en las sesiones de Geometría, tópicos que en general presentaban frases más largas y en algunos casos denotaban acciones y/o momentos particulares de la sesión. Estos fenómenos podrían sugerir que en las sesiones de Teoría de Números y Estadística la discusión sobre el problema se propició de forma más fructífera que en las de geometría. Por otro lado, no se encontró ningún patrón en los tópicos INTER que permitiera diferenciar entre Teoría de Números y Estadística.

1.3. Discusión

En este capítulo se ha presentado el desarrollo de un sistema que identifica diferentes tópicos en un conjunto de sesiones de RCP utilizando ASR y NLP. Además, el sistema entrega un ranking de las palabras más significativas de cada tópico, para poder dotarlos de interpretabilidad. También se ha mostrado la posibilidad del uso de este sistema para estudiar la variación de los tópicos a lo largo de las sesiones, generando visualizaciones interactivas y encontrando patrones que concuerdan con los cursos dictados en cada sesión. Por último, estos patrones observados fueron testeados estadísticamente encontrando diferencias significativas en la proporción de los tópicos para los distintos cursos.

Tras rankear las palabras en cada tópico usando tf-idf, fue posible encontrar tópicos que denotaban interacción entre los alumnos, con palabras como "creo", "pregunta", y '"entonces"; y tópicos asociados a un contenido en particular, con términos como "rectángulo" o "polígono". Más aun, fue posible identificar dos grupos preponderantes entre los tópicos de contenido, siendo estos geometría y números, e incluso inferir algunas acciones particulares asociadas a algunos tópicos de interacción. Adicionalmente, se observó que los tópicos estaban separados por la cantidad de palabras presentes en el vocabulario de RoBERTuito, encontrando dos clusters linealmente separables con el más pequeño conformado únicamente por frases de 5 palabras o menos.

Por su parte, el muestreo de las frases con mayor probabilidad de pertenencia a cada tópico reveló que las frases, particularmente en el cluster más grande, tenían una clara relación con el contenido asignado previamente a este. Sin embargo, ocasionalmente las frases en el cluster más pequeño parecían ser poco entendibles debido a su largo. A pesar de esto, se observó que las frases cortas con errores en la detección de voz fueron agrupadas en un sólo tópico, posiblemente debido al proceso de *tokenización* de RoBERTa. Así, estos análisis permitieron corroborar lo hipotetizado sólo desde las palabras claves de los tópicos.

Por último, tras estudiar las visualizaciones de las sesiones de Geometría y Teoría de Números, se pudo observar una clara preponderancia de los correspondientes tópicos de forma sistemática, mientras que las sesiones de Estadística presentaban mayoritariamente tópicos similares a las de Teoría de Números. Esto puede deberse principalmente a que los términos asociados a estadística quedaron mayoritariamente repartidos entre tópicos ELAB y NUM. Estas hipótesis fueron posteriormente confirmadas utilizando el test χ^2, el cual además reveló que las sesiones de Teoría de Números y Estadística presentaban mayores valores de tópicos INTER, lo cual podría sugerir una discusión más activa en estos grupos.

Estos resultados muestran que el sistema desarrollado permite dar una representación general de los tópicos presentes en el discurso de las sesiones y generar además una representación temporal de la sesión en función de estos. Además, se ha mostrado que esta representación de las sesiones puede captar diferencias entre distintas sesiones, como se observó con los distintos cursos dictados. Adicionalmente, como corolario de la metodología utilizada, se ha mostrado que con pocos datos anotados es posible transcribir automáticamente y con una calidad aceptable las sesiones de RCP de forma gratis y segura. Por ende, se espera que la metodología y los resultados mostrados en este capítulo puedan ser de utilidad a futuros investigadores interesados en obtener una representación general de sus sesiones de RCP.

Capítulo 2

Desarrollo de un Índice Temporal de Colaboración para Identificación de Interacciones Productivas

La toma de turnos a lo largo de una sesión de RCP determina la estructura temporal y social de las interacciones que la conforman [43]. Por un lado, esta representa el orden lógico y deductivo en que ciertas interacciones ocurren, como en el caso de responder a una pregunta hecha por un compañero [44]. Por otro lado, es posible identificar perfiles individuales de interacción de cada estudiante basado en las ocasiones en que a cada participante le es otorgado el turno y en la frecuencia con que esto ocurre [45]. Adicionalmente, estudios previos muestran que existen patrones de interacción, particularmente el tomar un turno ofrecido explícitamente a alguien más, que benefician a la equidad de participación del grupo y promueven una mayor cantidad de interacciones, factores que han sido identificados como claves para una colaboración exitosa [46].

A pesar de esto, la literatura también indica que una mayor cantidad de interacciones no basta para obtener una mejor solución al problema planteado, pues no todas las interacciones ayudan a avanzar en la formulación de esta [47]. En consecuencia, los momentos de interacción productiva, es decir, en los cuales se avanza en la resolución del problema, son acotados [48]. Debido a esto, es usual que codificadores humanos observen y anoten las sesiones de forma de identificar los momentos productivos a lo largo de una sesión. Sin embargo, esta metodología requiere de un equipo entrenado de codificadores y consume demasiado tiempo, por lo que no es posible escalarla a un gran número de sesiones de forma continua.

Por este motivo, el objetivo de este capítulo es proponer e implementar un índice temporal de colaboración que permita identificar los momentos de interacción productiva a lo largo de una sesión de RCP, evitando así la necesidad de anotar estos manualmente por un codificador humano. Particularmente, se proponen dos estadísticos, basados en la toma de turnos de los estudiantes, los cuales son luego aplicados sobre ventanas móviles a lo largo del tiempo obteniendo una curva para cada sesión. De esta forma, se espera que los máximos (mínimos) locales de la curva descrita por estos índices indiquen momentos de colaboración (no-)productiva.

Adicionalmente, una vez desarrollados los índices, se pidió a un codificador experto que

identificara los momentos de interacción productiva de un subconjunto de sesiones. Esto permitió comparar las curvas de los índices obtenidos con la codificación humana y reveló similitudes en los patrones de las curvas. Finalmente, tras la visualización de estas, se discretizó los índices fijando umbrales y se analizó el porcentaje de acuerdo con la codificación humana para distintos valores de los umbrales.

A continuación, la sección 2.1 presenta la metodología utilizada para la identificación de los turnos y la formulación de los índices propuestos. Posteriormente, en la sección 2.2 se explica la calibración de los índices propuestos, se analizan las curvas obtenidas, y se comparan los índices contra la codificación humana. Finalmente, la sección 2.3 presenta una breve discusión de los resultados y las implicancias de estos, y propone futuras lineas de investigación basadas en las limitaciones encontradas.

2.1. Metodología

Esta sección expone de forma detallada el procedimiento llevado a cabo para definir y calibrar un índice temporal de colaboración. Este proceso corresponde a tres fases a grandes rasgos: identificación de cambios de turno, cálculo de una batería de estadísticos desde los cambios de turno, y generación de un índice de colaboración desde los estadísticos calculados.

Tanto la identificación de cambios de turno, así como los estadísticos calculados y la formulación del índice, se basan en la idea de que los momentos de colaboración están caracterizados por un intercambio de ideas y una elaboracion y resolución de las mismas. Esto en principio daría origen a cambios de turno concentrados durante dicha interacción, lo cual se vería reflejado en los estadísticos calculados. Las subsecciones siguientes abarcan a cabalidad cada uno de los pasos mencionados anteriormente.

2.1.1. Identificación de los Cambios de Turno en el Discurso

La primera etapa para la formulación de índice de colaboración apunta a contestar la pregunta "¿quién habló a cada momento?" de forma automática, proceso conocido comúnmente como *diarización de voz* [49]. Esto permite analizar la dinámica de la sesión desde el punto de vista de la toma de turnos, motivado por la hipótesis de que los cambios de turno en el discurso durante la sesión de RCP contienen información acerca del desarrollo de la misma. Adicionalmente, dado que el proceso de diarización también es capaz de identificar distintas personas además de los cambios de turno, es posible estudiar también las dinamicas individuales de los estudiantes en términos de sus tiempos de intervención. Considerando estas características, se espera que utilizando de diarización de voz se pueda obtener información relevante para un índice temporal de colaboración.

Formalmente, un algoritmo de diarización de voz es un proceso que toma una señal de audio $x \in \mathcal{X}$ y retorna una serie de segmentos anotados $(t_k^i, t_k^f, s)_{k=1}^n \in ([0,T] \times [0,T] \times \{1,\ldots,\kappa(x)\})^*$, como se muestra en la figura 2.1, donde t_k^i, t_k^f son los tiempos de inicio y final del segmento, respectivamente; T es la duración de x en segundos y $\kappa(x)$ es el número total de hablantes en la señal. Para lograr esto, la tarea es dividida en una serie de pasos que incluye una segmentación inicial basada en detección de voz y de cambios de hablante, extracción de rasgos latentes de los segmentos, y agrupación de estos mediante clustering, los cuales se explican en detalle a continuación.

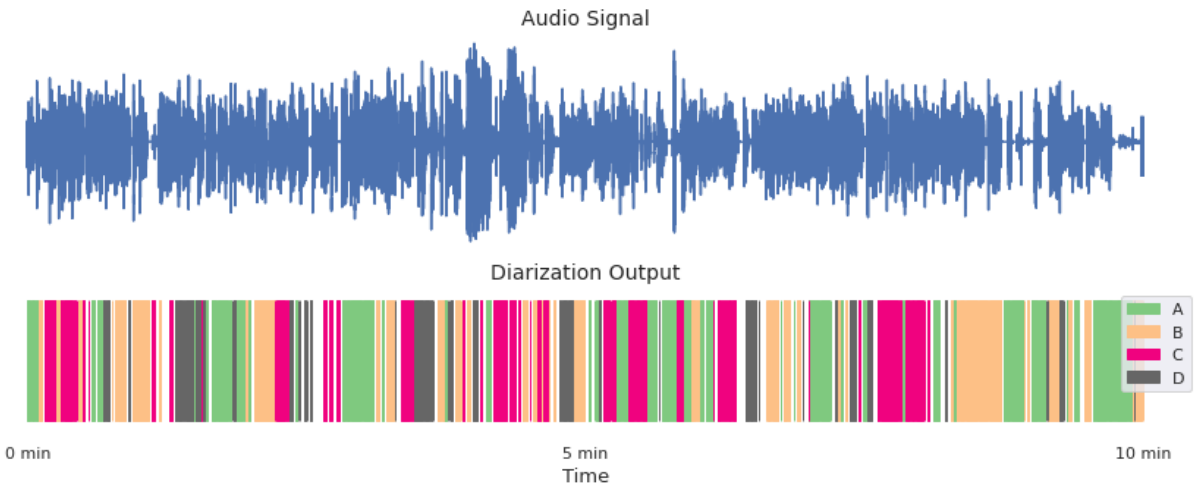


Figura 2.1: Ejemplo de diarización para 10 minutos de audio.

2.1.1.1. Segmentación Inicial

En general, la segmentación inicial en los sistemas de diarización de voz consta de la aplicación conjunta de dos algoritmos para obtener segmentos puros (es decir, con un único hablante). El primero identifica momentos donde efectivamente hubo discurso, tarea conocida como detección de voz o VAD (***V**oice **A**ctivity **D**etection*). Por otro lado, el segundo se encarga de encontrar los tiempos de cambio de turno a lo largo de la señal, tarea conocida como detección de cambios de hablantes o SCD (***S**peaker **C**hange **D**etection*).

Ambas tareas pueden ser formalizadas dentro del paradigma de etiquetado de secuencias [50], el cual se describe a continuación. Sea $x = (x_1, \ldots, x_T) \in \mathcal{X}, x_t \in \mathbb{R}^d$ una secuencia de variables extraída desde una señal de audio, usualmente aplicando una transformación sobre ventanas móviles de pocos milisegundos (e.g: espectros, amplitud, energía, etc.). Sea además $y \in \mathcal{Y}$ la secuencia de etiquetas $y = (y_1, \ldots, y_T), y_t \in \{0, \ldots, K-1\}$, donde K es el número de posibles etiquetas. Entonces, el objetivo del problema es aproximar la función $g : \mathcal{X} \longrightarrow \mathcal{Y}$ que asocia a cada secuencia de variables sus correspondientes etiquetas.

Usualmente, esta aproximación es lograda mediante una relajación del problema de etiquetado discreto en la cual se reemplaza el espacio de llegada de la secuencia $\{0, \ldots, K\}$ por el espacio Ω_K de las distribuciones de probabilidad K-discretas. Así, la función g es aproximada mediante una función $\hat{g} : \mathcal{X} \longrightarrow \Omega_K^*$, en la cual se busca minimizar alguna función de pérdida entre $\hat{y} = \hat{g}(x)$ e y, siendo la más común la entropía cruzada [51]:

$$L(y, \hat{y}) = -\frac{1}{T} \sum_{t=0}^{T} \sum_{k=0}^{K-1} 1_{\{y_t = k\}} \log(\hat{y}_t(k))$$

donde $\hat{y}_t(k)$ representa la probabilidad asociada a la clase k en tiempo t.

Sigue de lo anterior que ambos problemas son casos particulares del problema de etiquetado de secuencias con $K = 2$. Cabe notar que para este valor de K, $\hat{y}$ queda unícovamente

definida por cualquiera de sus componentes, por lo tanto, en adelante se referirá a $\hat{y}_t(1)$ simplemente como $\hat{y}_t$. Para el caso de VAD, basta considerar $y_t = 1$ si hubo discurso e $y_t = 0$ si no. Luego, se obtienen las probabilidades $\hat{y}_t$ aplicando la función aproximada $\hat{g}$ a x y se definen umbrales de entrada y salida (θ_{in} y θ_{out}) sobre $\hat{y}$, los cuales marcan el inicio y final de las regiones con voz.

De forma similar, para el problema de SCD basta con considerar $y_t = 1$ si hubo un cambio de hablante en t e $y_t = 0$ si no. Posteriormente, se definen un largo de intervalo l y un umbral θ_{peak} de manera que cada para t, el conjunto de máximos locales del intervalo centrado en t y de largo l con probabilidad mayor que θ_{peak} son marcados como cambios de turno. Así, el algoritmo de segmentación inicial identifica en primera instancia segmentos con actividad de voz utilizando VAD, y posteriormente divide estos segmentos en los tiempos marcados como cambios de turno para obtener segmentos puros.

2.1.1.2. Extracción de Rasgos Latentes de los Segmentos

Una vez obtenidos los segmentos puros, el siguiente paso en un algoritmo de diarización de voz consta de identificar qué segmento corresponde a cada hablante. Es decir, queremos aproximar la función $h : \mathcal{X} \longrightarrow \mathbb{N}$ que a un segmento puro $\tilde{x} \in \mathcal{X}$, subsecuencia de una señal x, le asigna un hablante $k \leq \kappa(x)$. Dada la complejidad del problema discreto, este suele aproximarse en dos pasos: primero se obtienen representaciones latentes de las cualidades acústicas de cada hablante, para luego agrupar estas mismas en el espacio usando algún algoritmo de *clustering*.

A continuación, se presenta la formalización del problema de representaciones latentes para identificación de hablantes [52]. Sean un conjunto de segmentos puros $X = (x_j)_{j=1}^n$ y una (*pseudo-*)distancia d en $\mathbb{R}^D$ (comunmente la similaridad de coseno). Definimos por $\mathcal{T} = \{(x_a, x_p, x_n) \in X^3 : h(x_a) = h(x_p), h(x_a) \neq h(x_n)\}$ al conjunto de todas las combinaciones de pares positivos (x_a, x_p) y negativos (x_a, x_n) que se pueden formar a partir de algún $x_a \in X$. Entonces el objetivo es encontrar una función de representación $f : \mathcal{X} \longrightarrow \mathbb{R}^D$ que cumpla (o se acerque lo máximo posible bajo alguna noción a definir) la siguiente propiedad para todo $\tau = (x_a, x_p, x_n) \in \mathcal{T}$:

$$d(f(x_a), f(x_p)) < d(f(x_a), f(x_n))$$

Es decir, dos representaciones que correspondan al mismo hablante deberían siempre estar más cerca entre sí que dos representaciones de hablantes distintos. En la práctica, el estado del arte usualmente es modelar la función f mediante redes neuronales recurrentes, utilizando una función de pérdida motivada por la propiedad deseable anterior, llamada pérdida de tripletes [53]:

$$\mathcal{L}(\mathcal{T}) = \sum_{\tau \in \mathcal{T}} \max(0, \alpha + \Delta_\tau), \;\; \Delta_\tau := d(f(x_a), f(x_p))^2 - d(f(x_a), f(x_n))^2$$

Donde $\alpha > 0$ es un hiperparámetro que controla qué tan fuerte debería ser la separación espacial entre hablantes llamado *margen de seguridad*.

Dado que calcular las diferencias de todos los tripletes posibles no es eficiente en tiempo ni memoria, y considerando que sólo los tripletes que cumplen $\Delta_\tau + \alpha > 0$ aportan a la función de pérdida, es común usar estrategias de muestreo para aproximar la función de

pérdida en cada iteración de la red neuronal, siendo la más común la estrategia de muestreo negativo [54]. En esta estrategia, por cada hablante k se extraen n muestras, dando origen a $\kappa(X)n(n-1)$ pares positivos. Luego, para cada uno de estos pares, se elige x_n de entre los otros $(\kappa(X)-1)n$ segmentos muestreados, de forma que se maximice Δ_τ. Finalmente, se busca minimizar esta aproximación de $\mathcal{L}(\mathcal{T})$ en cada iteración de la red neuronal de forma de separar lo mejor posible los hablantes en el espacio.

La implementación del modelo de diarización fue hecha utilizando la librería `pyannote.audio`, la cual provee modelos preentrenados y ya acoplados para todos los pasos necesarios mencionados en esta subsección y la siguiente [50].

2.1.1.3. Clustering de las Representaciones Latentes

El proceso de *clustering* forma una parte fundamental de cualquier sistema de diarización de voz, al ser el mecanismo mediante el cual se asignan las etiquetas de los distintos hablantes a cada segmento [49]. Actualmente, existen múltiples métodos de clustering, que a grandes rasgos se pueden clasificar en 4 tipos según la noción en la que se basan para agrupar puntos: jerárquicos, por centroides, por distribuciones, y por densidad [55]. Cabe destacar que en nuestra versión del problema, el número de hablantes en una señal es también una variable a estimar. Esto imposibilita el uso de métodos de clustering por centroides o distribuciones, los cuales necesitan conocer de antemano el número de grupos a modelar.

Considerando esto, se decidió utilizar el método de *propagación por afinidad*, el cual a la fecha de escritura constituye el estado del arte para diarización de voz con estimación de número de hablantes [56], [21]. Este algoritmo, a diferencia de otros con inicialización aleatoria, considera que todos los puntos son posibles representantes de un cluster. Luego, la idea general consta de intercambiar mensajes entre los puntos respecto de quién es mejor candidato para representar a cada punto, basado en la similaridad con respecto a dicho candidato, y en cómo los otros puntos consideran a este.

Formalmente, sean $x_1, \ldots, x_n \in \mathbb{R}^d$, y $s : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ una función de similaridad, usualmente considerada como el inverso aditivo de la distancia angular o euclidiana al cuadrado, y definimos $S = (s_{ij})_{i \neq j},\ s_{ij} = s(x_i, x_j)$. Los valores en la diagonal s_{ii} son considerados un hiperparámetro del modello llamado el vector de *preferencias* sobre los vectores x_i y controla la prioridad que tiene cada punto para ser elegido un como un representante.

Adicionalmente, definimos las matrices auxiliares de *responsabilidad* y *disponibilidad*, denotadas por $R = r_{ij}$ y $A = a_{ij}$ respectivamente, las cuales se encargan de almacenar y actualizar la información intercambiada por los puntos. A grandes rasgos, r_{ij} representa la preferencia que tiene el punto x_i por x_j como representante, comparado con otros posibles representantes, mientras que a_{ij} representa cuan importante es x_j como representante para x_i, pero considerando las responsabilidades que le otorgan los demás puntos como representante.

Por último, para las condiciones de término del algoritmo, se utiliza la matriz de criterio $C = A + R$, en conjunto con un número máximo de iteraciones M y un parámetro de paciencia p. El representante de cada punto x_i en cada iteración está dado por $x_{\tilde{j}}$, donde $\tilde{j} \in argmax(\{c_{ij} : j \in 1, \ldots, n\})$. Así, el algoritmo recibe como inputs S, M y p, y comienza inicializando las matrices A y R como zeros.

En cada iteración, primero es actualizada la matriz de responsabilidades mediante la regla

$$r_{ij} \leftarrow s_{ij} - \max_{j' \neq j}(a_{ij'} + s_{ij'})$$

Es decir, la responsabilidad de i a j es actualizada mediante la diferencia entre la similaridad i, j, y el máximo de la suma de las similaridades y las disponibilidades hacia otros puntos. Posteriormente, se actualiza la matriz de disponibilidades mediante la regla

$$a_{ij} \leftarrow \begin{cases} \min\left(0, r_{jj} + \sum_{i' \notin \{i,j\}} \max(0, r(i', j))\right), & \text{si } i \neq j \\ \sum_{i' \neq j} \max(0, r(i', j)), & \text{si } i = j \end{cases}$$

Las disponibilidades cruzadas, desde j hacia i, son actualizadas mediante la suma de la auto-responsabilidad, y las responsabilidades positivas que j tiene hacia otros puntos (distintos de i); mientras que las auto-disponibilidades se actualizan mediante la suma de las responsabilidades positivas hacia otros puntos.

Cabe destacar que se consideran sólo las responsabilidades positivas con otros puntos, dado que sólo es necesario que un ejemplar explique bien *algunos* de los puntos, sin importar qué tanto falla en los demás. Así, el algoritmo asigna en cada iteración los representantes de cada punto mediante la matriz de criterios, y el algoritmo termina cuando los representantes no cambian durante p iteraciones consecutivas, o al alcanzar M iteraciones. De esta forma, a cada segmento le es asignado un representante basado en sus rasgos latentes.

2.1.2. Formulación de un Índice Temporal de Colaboración

Una vez se obtuvo los turnos en las grabaciones y se asignaron a los distintos hablantes, se buscó estudiar una posible relación entre los cambios de turno y los momentos de colaboración a lo largo de la sesión. Para esto, se comenzó por realizar un análisis exploratorio de las anotaciones obtenidas mediante el proceso de diarización de voz y se estudió la cantidad local de hablantes a lo largo de las sesiones.

Posteriormente, esto motivó la idea de aplicar una versión localizada de los estadísticos anteriormente calculados y combinarlos para formular un índice que permita identificar momentos de colaboración a lo largo de las sesiones. Finalmente, se aprovechó un subconjunto de sesiones previamente anotadas de forma manual con una rúbrica de colaboración para comparar el índice desarrollado con la codificación de la rúbrica. A continuación se detalla la metodología utilizada para la formulación del índice.

2.1.2.1. Análisis Exploratorio y limpieza de las Anotaciones

En primer lugar, para tener una mirada general de cuánto hablan los alumnos a lo largo de la sesión, se comenzó por estudiar la duración de la sesiones y el tiempo total de discurso en cada una, así como la proporción entre estas variables y el número de hablantes encontrados en cada sesión. Esto reveló que en más del 80 % de las sesiones se habló al menos la mitad del tiempo y el 40 % mostró haber hablado más del 80 % del mismo. Esto incitó a estudiar más detenidamente las sesiones con poco tiempo de habla.

Adicionalmente, el número de hablantes encontrados utilizando el sistema de diarización

de voz presentó valores entre 2 y 12 (mínimos con una grabación cada uno) con moda de 6. Esto motivó a estudiar las grabaciónes con más de 7 hablantes encontrados, dado que el máximo presente en las sesiones, según los profesores que las llevaron a cabo, era de 6 alumnos (7 hablantes contando al profesor). La figura 2.2 presenta las distribuciones de la duración de la grabación, el tiempo neto hablado, y el total de hablantes encontrados.

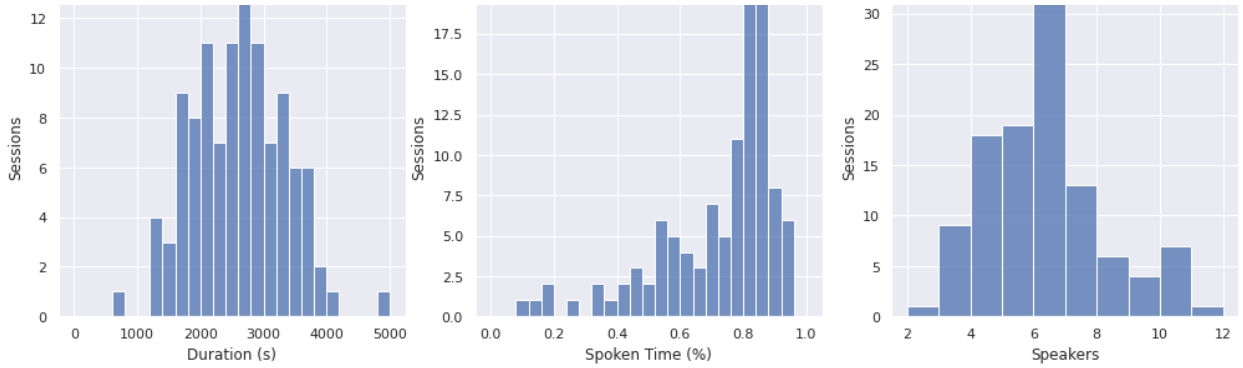


Figura 2.2: De izquierda a derecha, histogramas para la duración de la grabación, el total de tiempo hablado por sesión y el número de hablantes encontrados en cada grabación.

Durante el análsis de las grabaciones con poco tiempo de habla o muy corta duración, se encontró una grabación de menos de 15 minutos y dos grabaciones con menos del 5 minutos netos de tiempo de habla, las cuales se removieron del conjunto de datos tras ser analizadas. Por otro lado, el estudio de las grabaciones con muchos hablantes reveló que en general el número de estos podía verse inflado por falsos positivos en el paso de detección de voz, principalmente por ruidos u otras voces de fondo, los cuales posteriormente en el proceso de propagación por afinidad eran agrupados aparte, dada la diferencia de las cualidades acústicas con respecto al discurso normal.

Dado que estos segmentos en general aparecían de forma remota y aislada en las grabaciones, y con duraciones muy cortas, se decidió fijar un umbral mínimo sobre la proporción del total de tiempo hablado para considerar a un hablante como válido dentro de la grabación. Tras estudiar el histograma del mínimo de tiempo que algún hablante tuvo durante la sesión, se encontró una moda muy superior en el intervalo $[0, 0.01]$, teniendo 18 sesiones, mientras que el máximo entre todos los demás intervalos fue de 10.

Considerando que este valor se alcanzaba en el intervalo $[0.06, 0.07)$, valor razonable para el umbral, se decidió fijar el umbral en 0.06. La figura 2.3 presenta los histogramas del tiempo mínimo y el número de hablantes antes y después de aplicar el filtro. Se puede observar como el número de hablantes se concentra mayoritariamente en el intervalo $[0, 7]$ tras la limpieza de los datos, mientras que el porcentaje de tiempo mínimo se distribuye de forma proporcional a lo encontrado antes de la limpieza entre los intervalos mayores a 0.06

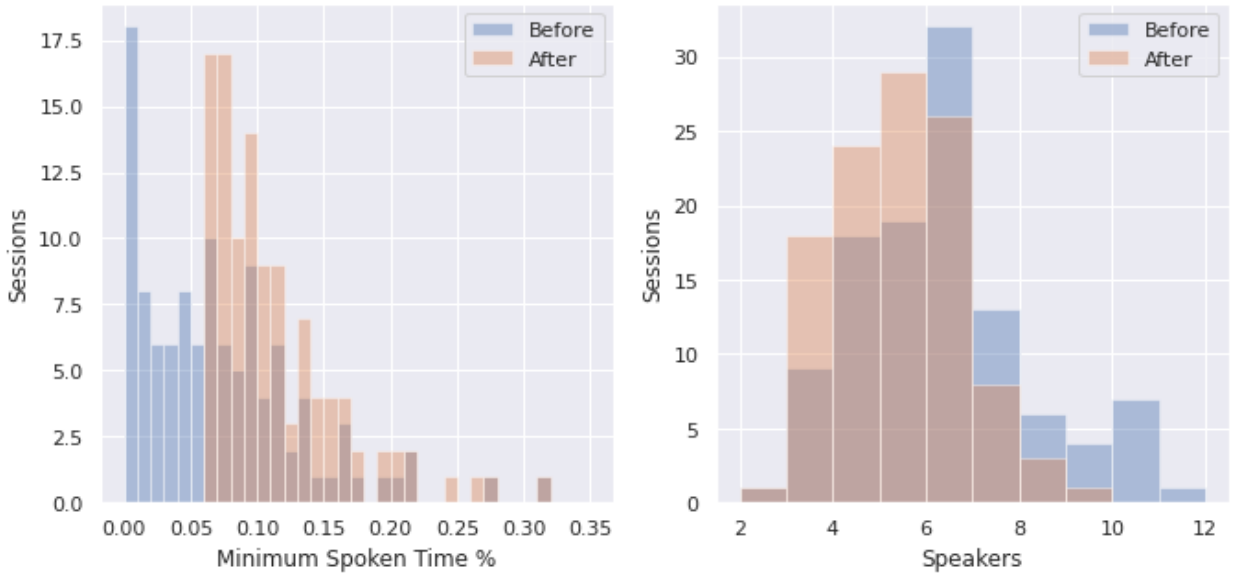


Figura 2.3: Histogramas para la mínima proporción de tiempo hablado (izquierda) y el número de hablantes (derecha) encontrados en cada grabación.

En segundo lugar, para comenzar a explorar aspectos relativos a la toma de turnos, se calcularon los porcentajes de tiempo que cada alumno habló a lo largo de la sesión, con respecto al total de tiempo hablado y sobre la cantidad de hablantes válidos. Luego, se calculó una batería de estadísticos con respecto a las anotaciones y se estudiaron los correspondientes histogramas de forma similar a la sección anterior.

Los estadísticos calculados fueron el porcentaje de tiempo máximo y mínimo de habla entre los alumnos de la sesión, el número de turnos distintos a lo largo de la sesión, el número de turnos por unidad de tiempo, los largos promedio, mínimo y máximo de los turnos; y la distancia de Hellinger entre la distribución de los tiempos de habla y la distribución uniforme sobre el número de alumnos en la sesión.

Posteriormente, se analizaron en detalle las grabaciones que presentaron valores extremos y aislados en los histogramas de los estadísticos calculados. Sin embargo, se determinó que estas correspondían efectivamente a instancias de RCP por lo que se mantuvieron en el conjunto de datos. Por último, se generaron visualizaciones de las diarizaciones obtenidas en una linea de tiempo. La figura 2.4 muestra un ejemplo de las visualizaciones obtenidas.

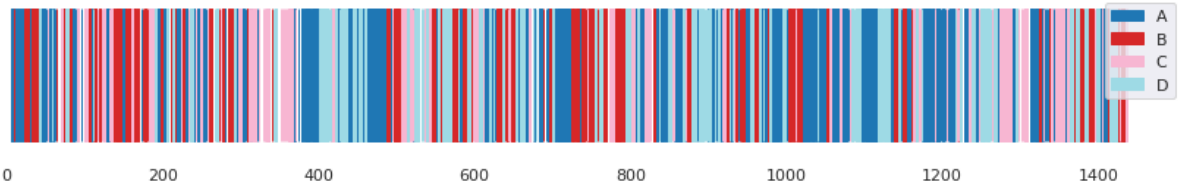


Figura 2.4: Ejemplo de diarización obtenida para una sesión de RCP.

2.1.2.2. Formulación y Calibración del Índice

Dado que la colaboración no se da de forma uniforme a lo largo de la sesión, es necesario estudiarla a lo largo de la dimensión temporal. Con esto en mente, se propuso localizar el análisis aplicando ventanas móviles a lo largo del eje tiempo, con largo 2δ y paso h parametrizables. Esto permite analizar la evolución temporal de la toma de turnos, a la vez que se puede controlar la suavidad de la curva ($h \longrightarrow 0$) y el tamaño de la vecindad a estudiar (δ).

En particular se realizaron dos experimentos, correspondientes a dos propuestas para un índice unidimensional de colaboración, motivadas de las observaciones generadas durante el análisis de los estadísticos globales, a los cuales se denominó *índices automatizados*. En primer lugar, se comenzó por definir una noción de hablante activo sobre una ventana de tiempo, de forma similar a lo realizado para encontrar a los hablantes válidos para una clase. Es decir, se definió un umbral θ tal que para cada ventana de tiempo $[t_i, t_{i+1})$, un hablante s se consideró activo durante esa ventana si la intersección (en tiempo) entre los segmentos puros correspondientes a s y dicha ventana era de al menos $\theta(t_{i+1} - t_i)$ segundos. Este estadístico corresponde al primer índice automatizado.

Posteriormente, se calculó una batería de estadísticos similar a la aplicada globalmente y se estudiaron las distribuciones de estos. Los estadísticos calculados comprenden el porcentaje máximo de tiempo hablado durante la ventana por algún alumno, el número de turnos distintos encontrados, el largo promedio de estos, el número de hablantes activos durante la ventana, y la proporción de tiempo hablada durante la ventana. En general, los estadísticos presentaron distribuciones similares a una normal, con excepción del tiempo hablado, el cual presentó una distribución similar al caso global, concentrando la mayoría de los segmentos en el intervalo $[0.7, 0.9)$.

Dada la cantidad de estadísticos calculados, para facilitar posteriores análisis se decidió colapsar la batería en un índice unidimensional, esto se hizo utilizando PCA (***P**rincipal **C**omponent **A**nalysis*) [57]. Esta técnica permite proyectar variables $x \in \mathbb{R}^n$ conservando la mayor proporción de varianza explicada. Esto se logra encontrando iterativamente una dirección ortonormal a las anteriores, w_i que minimize la distancia entre los puntos en $\mathbb{R}^n$ y su proyección en el espacio $span(\{w_1, \ldots, w_i\})$. De esta forma, para proyectar en un espacio $d-$dimensional ($0 < d < n$), basta con aplicar a los vectores x la transformación lineal que tiene por columnas $w_1, \ldots, w_d$. En este caso, dado el objetivo de obtener una sola dimensión, se decidió utilizar la primera componente principal.

2.2. Análisis de Resultados

A continuación se presentan los resultados observados durante la formulación e implementación de los índices automatizados. Primero, se comienza por seleccionar un valor apropiado para el largo de ventana analizando las curvas de hablantes activos obtenidas para los distintos valores de este. Posteriormente, se estudia la composición de las 2 primeras componentes encontradas tras aplicar PCA a la batería local de estadísticos, es decir, los pesos asociados y los porcentajes de varianza explicados. Luego, se describen y comparan los patrones encontrados utilizando los dos índices automatizados. Por último, se describe un experimento realizado para comparar los índices automatizados a la codificación manual de un experto

sobre un subconjunto de 9 clases, y se presentan los resultados de este.

2.2.1. Selección del Largo de Ventana

En primer lugar, es importante estimar un valor para el largo de ventana (l) de forma de balancear las propiedades locales y globales en los índices calculados. Es decir, se busca que la ventana de contexto sea suficientemente amplia para que las curvas presenten intervalos claros de crecimiento y decrecimiento, a la vez que se desea evitar que se pierda dicha estructura local debido a valores demasiado altos de l. Para esto, se estudiaron las curvas obtenidas para valores de l en $\{30, 60, 180, 300, 500\}s$ con paso $h = l/2$ buscando las propiedades anteriormente descritas. La figura 2.5 muestra ejemplos de las curvas obtenidas sobre dos sesiones.

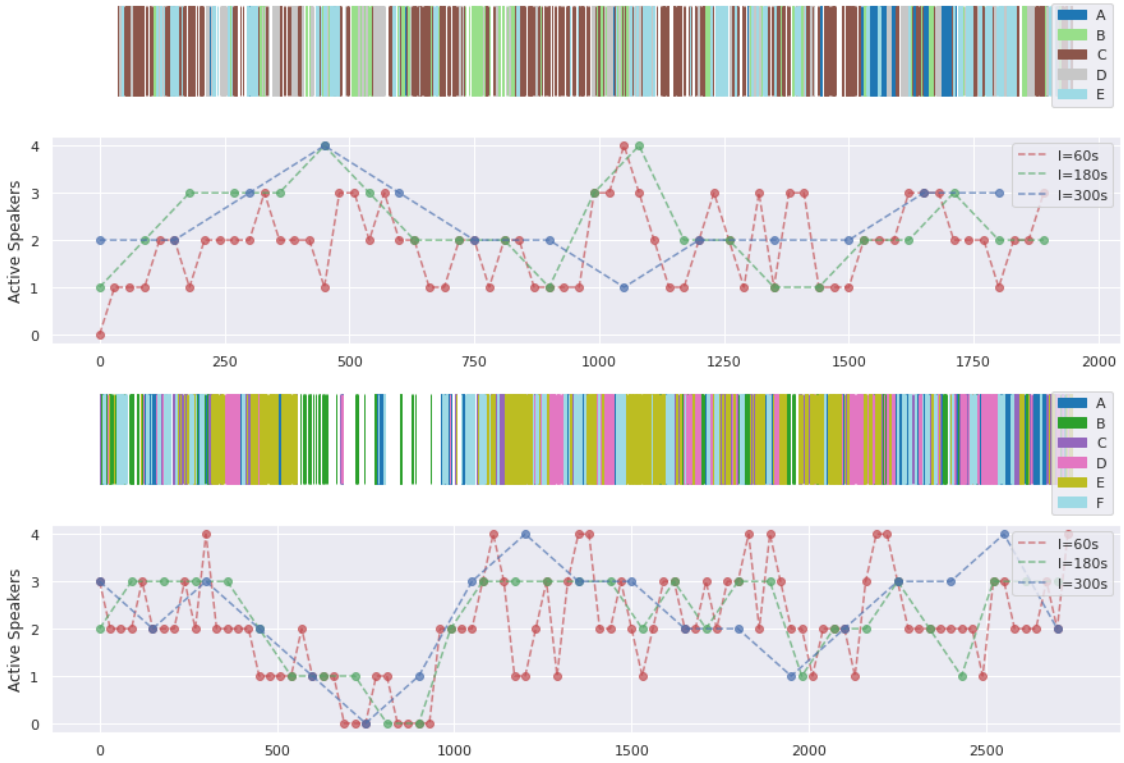


Figura 2.5: Número de hablantes activos a lo largo del tiempo para distintos valores del largo de ventana (l) en dos sesiones. En ambos casos se observa que para $l = 60s$ y $l = 300s$ se priorizan excesivamente las estructuras local y global, respectivamente.

En general, se observó que para valores de $l \leq 60s$, el número de hablantes estaba demasiado localizado, pues oscilaba constanstemente sin mostrar intervalos claros de crecimiento o decrecimiento. Por otro lado, para valores de $l \geq 300s$, las curvas conservaban poca estructura local y presentaban muy poca variación, lo cual dificultaría los análisis posteriores. Adicionalmente, valores muy altos de l reducían considerablemente el número de puntos totales en la curva en este experimento, dada la elección de h. Con esto en mente, se decidió que un balance razonable entre estructura local y global en los estadísticos se alcanzaba en $l = 180s$. Esto permitirá posteriormente comparar patrones entre las distintas curvas en términos de sus intervalos de crecimiento y de sus mínimos y máximos locales.

2.2.2. Estudio de las Componentes Principales Encontradas

Dado que la técnica de componentes principales genera una proyección lineal sobre un espacio de menor dimensión, una de sus ventajas consiste en la interpretabilidad de sus coeficientes. Adicionalmente, puesto que en cada iteración se minimiza la distancia entre la proyección y el vector original, es posible estudiar los porcentajes de varianza explicada por cada nuevo eje agregado en cada iteración. Considerando esto, se decidió explorar las relaciones en estos coeficientes antes de estudiar el comportamiento del índice generado.

En primer lugar, se comenzó por estandarizar las variables a media 0 y varianza unitaria, para evitar una desproporción en los pesos y los porcentajes de varianza explicados debido a las diferencias entre los rangos de los estadísticos. Una vez estandarizadas las variabes, se aplicó PCA sobre estas y se estudiaron los coeficientes de las dos primeras componentes principales obtenidas y los porcentajes de varianza explicados por estos.

Los resultados revelan que la primera componente principal explica un 54.3 % de la varianza de los datos en el espacio original, mientras que la segunda explica un 32.4 %. Particularmente, el valor asociado a la primera componente sugiere que este captura gran parte de la información contenida en la batería de estadísticos, considerando el número inicial de estos. Por otro lado, se observó que las componentes obtenidas tras aplicar PCA correlacionaban negativamente con los estadísticos calculados, particularmente en la primera componente.

Dado que la intuición indica que varios de los estadísticos calculados deberían aportar a la colaboración (como el número de segmentos y el porcentaje de tiempo hablado, por ejemplo), se decidió rotar las componentes encontradas cambiándole el signo a la matriz de pesos. Esto permite obtener un índice cuyo signo está alineado con la idea intuitiva de colaboración detrás de los estadísticos calculados.

Tras aplicar la rotación, se estudió los pesos asociados a cada índice, encontrando que el mayor y menor aporte a la primera componente correspondían al porcentaje de tiempo hablado, y el largo promedio de los segmentos, con un coeficiente de 0.59 y 0.16, respectivamente. La tabla 2.1 presenta los coeficientes obtenidos tras la rotación y los porcentajes de varianza explicados, mientras que la figura 2.6 presenta el círculo de correlaciones sobre las dos componentes calculadas.

Tabla 2.1: Coeficientes y porcentajes de varianza explicados obtenidos tras aplicar PCA a la batería de estadísticos estandarizada. De izquierda a derecha: TM: proporción de tiempo máxima hablada por un sólo alumno, NT: número de turnos encontrados en la ventana, HA: número de hablantes activos en la ventana, LP: largo promedio de los segmentos en la ventana, TH: proporción de tiempo hablado entre todos los alumnos en conjunto, VE: varianza explicada por la componente.

Componente	TM	NT	HA	LP	TH	**VE**
PCA 1	0.40	0.46	0.50	0.16	0.59	54.4 %
PCA 2	-0.50	0.43	0.30	0.68	0.8	32.4 %

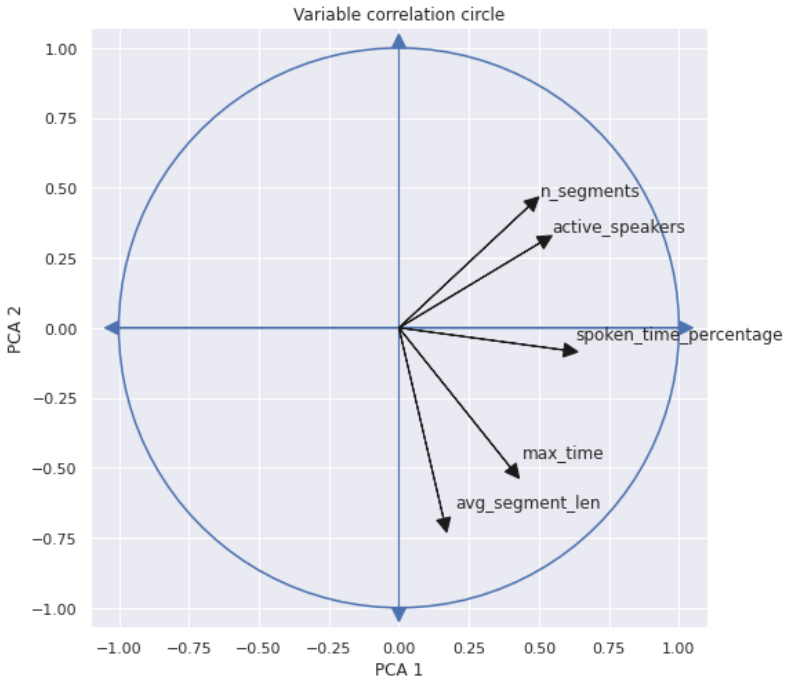


Figura 2.6: Círculo de correlaciones entre las dos primeras componentes principales y los estadísticos calculados.

2.2.3. Análisis Temporal de los Índices Propuestos

Tras haber estudiado las componentes obtenidas y la distribución de los estadísticos locales, se procedió a estudiar los patrones que se presentaban temporalmente en los dos índices calculados. Esto se hizo a través de la visualización de las curvas que ambas variables describían a lo largo del tiempo en conjunto con las anotaciones de diarización obtenidas.

En primer lugar, se comenzó por estudiar el efecto de fijar distintos valores para el umbral (θ) de los hablantes activos. Para esto, se graficó el número de hablantes activos para $\theta \in \{0, 0.1, 0.2, 0.3\}$ y se estudiaron las curvas obtenidas. En general, se encontró que para $\theta = 0$, las curvas se mantenían de forma constante en el máximo de hablantes activos, mientras que para valores de $\theta \geq 0.3$, las curvas rara vez sobrepasaban los dos hablantes activos. La figura 2.7 muestra ejemplos de esto para dos clases.

Por otro lado, para $\theta \in [0.1, 0.2]$, las curvas presentaban mayor oscilación, lo cual permite enriquecer el posterior análisis de los patrones que se presentan en estas. Dado esto, y considerando la distribución del número de hablantes válidos del análisis global, se decidió fijar el valor de θ en 0.15. Cabe mencionar que debido a la posibilidad de encontrar varios alumnos interviniendo activamente en la conversación en una ventana corta de tiempo, se descartaron valores para $\theta > 0.3$, pues esto acotaría superiormente el número de hablantes activos a 3 o menos.

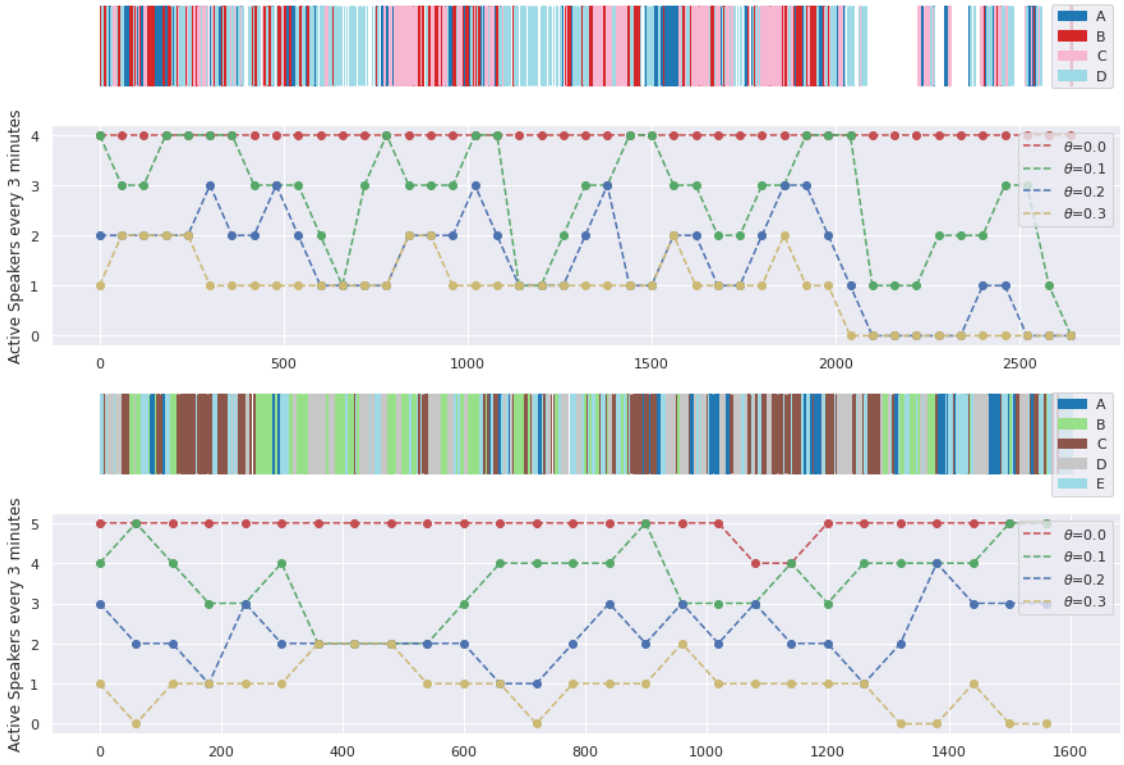
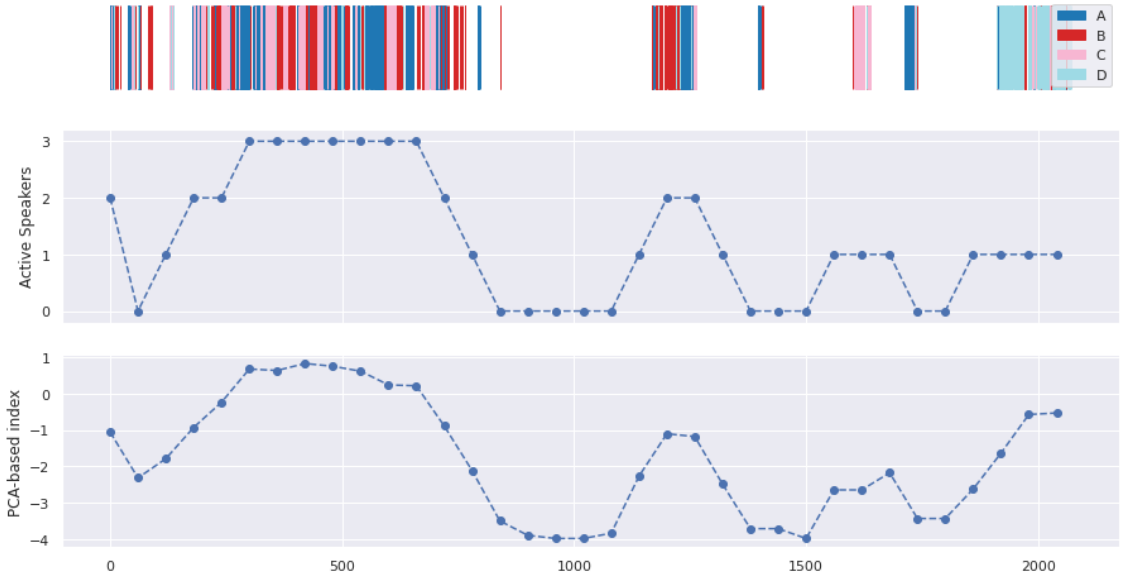


Figura 2.7: Número de hablantes activos a lo largo del tiempo para distintos valores del umbral (θ) en dos sesiones, con ventanas de 3 minutos y paso de 1.

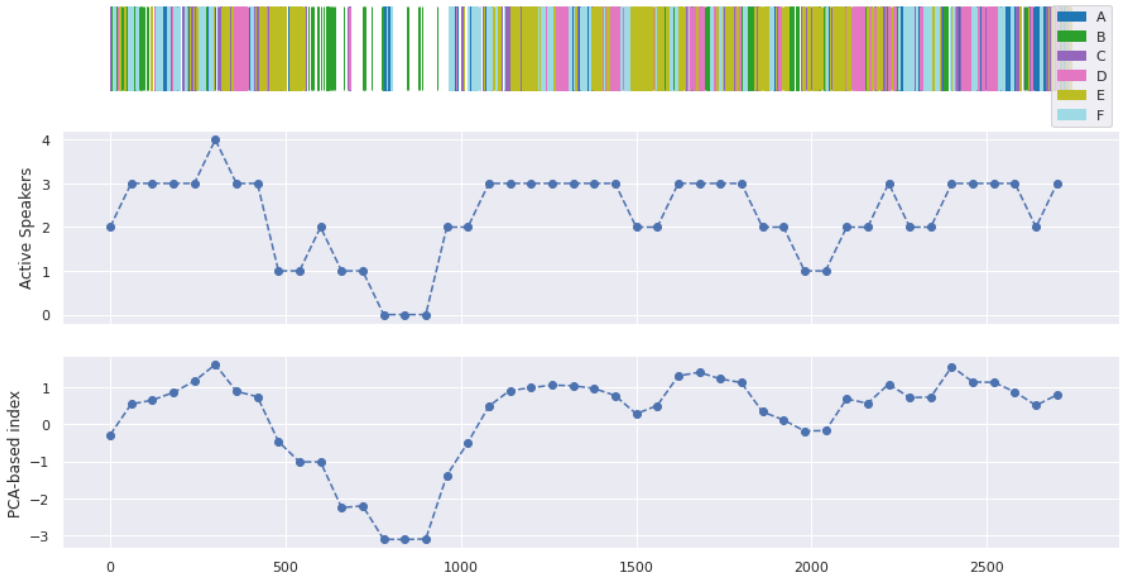
Posteriormente, se procedió a comparar los comportamientos observados en ambos índices entre sí y con respecto a la diarización obtenida. En general, ambas curvas presentaron similitud en términos de los intervalos de crecimiento y decrecimiento. Naturalmente, esto implicó también una similitud entre la posición temporal de los mínimos y máximos locales de estas. Esto se debe en gran parte al peso de la cantidad de hablantes activos en la primera componente, siendo este el segundo estadístico de la batería que más aporte a dicha componente.

Más aún, a grandes rasgos, para ambos se pudo observar dos patrones mayoritarios entre las distintas sesiones. El primero correspondía a un intervalo con valores altos en ambos indicadores al principio, seguido de valores bajos con eventuales picos en intervalos relativamente cortos. Esto puede corresponder a una inicial discusión y comprensión del problema por parte del grupo completo, para continuar trabajo personal y discusión/ resolución de dudas esporádicas. Un ejemplo de estas sesiones puede verse en la figura 2.8.a.

Por el contrario, el segundo patrón encontrado mostraba una oscilación constante en el número de hablantes activos a lo largo del tiempo, sin presentar momentos claros de mayor interacción que otros entre los estudiantes, pero presentando valles en los índices en momentos particulares, principalmente debido a momentos de silencio o donde habló una sola persona. Esto se puede observar en la figura 2.8.b. Este patrón puede deberse a una discusión más bien constante a lo largo del tiempo con un enfoque más grupal y menos trabajo individual.



(a) Sesión 1. Se observa una discusión inicial larga marcada por los valores altos en los índices, seguida de interacciones cortas marcadas por picos en ambas curvas.



(b) Sesión 2. Se observa una discusión constante a lo largo del tiempo dada por una oscilación de ambos índices alrededor de 2-3 y 0 respectivamente, con un valle entre los 12 y los 15 minutos debido al silencio.

Figura 2.8: Número de hablantes activos e índice basado en PCA a lo largo del tiempo para dos sesiones representativas de los patrones encontrados.

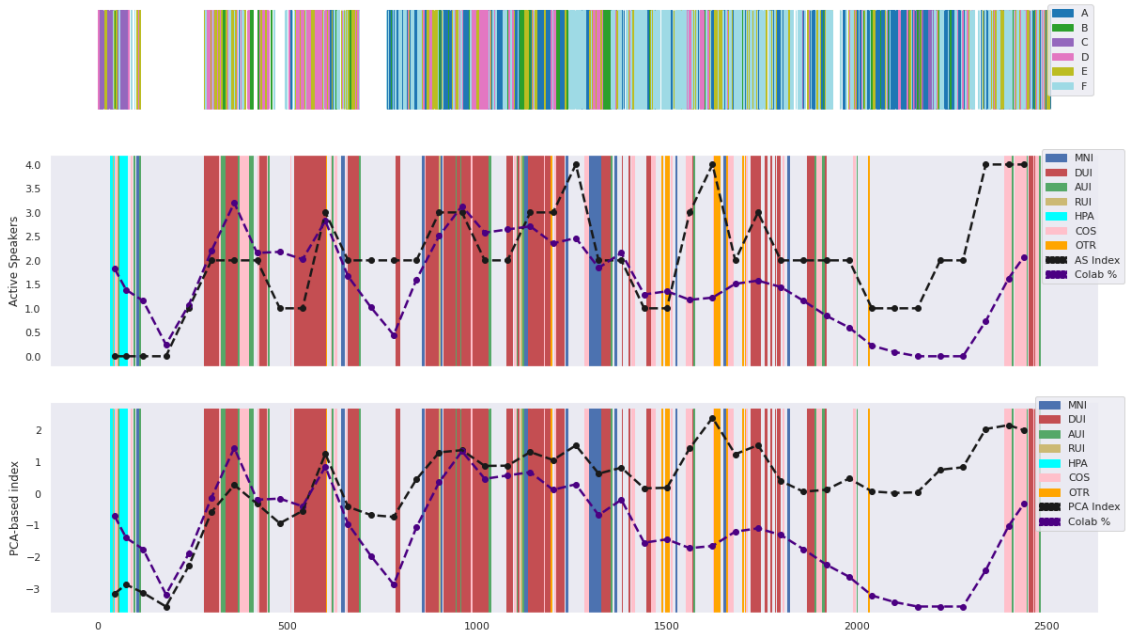
2.2.4. Comparación Entre los Índices Propuestos y la Codificación Humana

El último experimento llevado a cabo en esta tésis consistió en comparar los índices automatizado con respecto a la codificación manual de un experto, para obtener una respuesta a la pregunta de investigación de este capítulo. Para esto, se aprovecharon las anotaciones solicitadas a los investigadores del proyecto FONDEF.

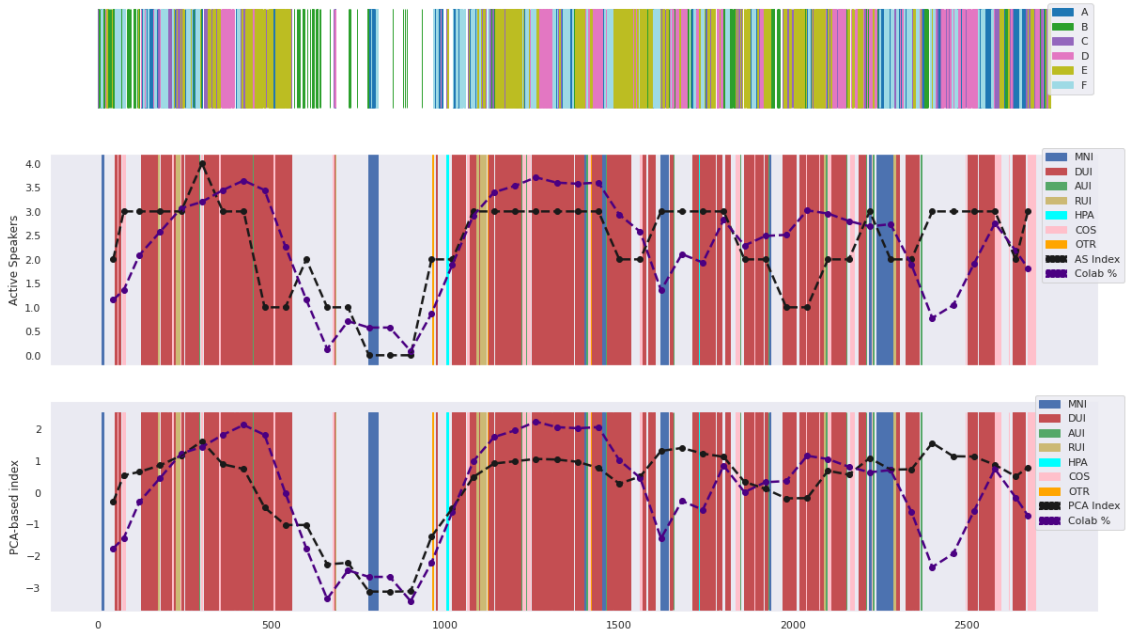
En primer lugar fue necesario adaptar la resolución temporal de la codificación manual para poder compararla con los índices propuestos, dado que las codificaciones manuales venían dadas por intervalos de largo variable, mientras que los índices se calculan sobre ventanas con largo y paso definidos. Esto se realizó mediante un acercamiento similar al utilizado para medir los hablantes activos a lo largo de la ventana. Particularmente, para cada ventana fija se calculó la proporción de tiempo que esta intersectaba a los segmentos codificados. Cabe mencionar que esta proporción tiene por cota superior 1, pues las distintas acciones en la rúbrica son consideradas mutuamente excluyentes. A este índice se le denominó *índice manual*

Una vez calculado el índice manual para cada ventana, esta se escaló de forma de poder visualizarlo comparativamente contra ambos índices automatizados. En el caso del número de hablantes activos, el índice manual fue multiplicado por el número máximo de hablantes presentes en una ventana a lo largo de dicha sesión. Por otro lado, dado que las componentes principales fueron calculadas considerando todas las sesiones, en este caso se aplicó una transformación lineal al índice manual con tal de escalarlo desde el intervalo $[0, 1]$ en $[PCA_{min}, PCA_{max}]$, donde PCA_{min}, PCA_{max} son los valores mínimo y máximo encontrados para la primera componente principal a lo largo de todas las sesiones, respectivamente.

En general, se pudo observar similaridad en los patrones de las curvas, particularmente en términos de los intervalos crecimiento y decrecimiento y las posiciones de los mínimos y máximos. Sin embargo, muchas veces las curvas parecían desplazadas una con respecto a la otra, estando los índices automatizados por sobre el manual, y ocasionalmente las proporciones entre los mínimos y máximos locales parecían realmente exageradas entre ambas curvas. Adicionalmente, se encontró máximos locales en los índices automatizados que no se correspondían con ningún máximo en la curva del índice manual. Estos fenómenos pueden deberse a que los índices propuestos no tienen capacidad para discernir entre frases que aporten y que no aporten a la resolución del problema, sino que sólo son capaces de identificar la presencia de una frase y diferenciar quien la dijo. La figura 2.9 muestra el índice manual y los índices automátizados para dos sesiones.



(a) Sesión 1. Se observa una similitud marcada, particularmente en los primeros dos tercios de la sesión, encontrando falsos positivos al rededor de los 700 y los 1600 segundos, además de una diferencia en la tendencia hacia el final de la sesión.



(b) Sesión 2. Se observa similitud a entre los patrones de las curvas a lo largo de todo el tiempo, encontrando falsos positivos al rededor de los 1700 y los 2400 segundos.

Figura 2.9: Índice manual versus índices automatizados para dos sesiones.

Dado que desde el análisis visual fue complicado determinar la calidad de los índices propuestos con respecto a la codificación humana, y considerando que la clasificación de la rúbrica es discreta, se decidió abarcar el problema como uno de clasificación, de forma de obtener una métrica para evaluar el desempeño de los índices automatizados con respecto de la clasificación humana. Para esto, se definió un umbral de forma de binarizar el índice manual balanceadamente. En particular, se estudió la distribución de este, encontrando una

mediana cercana al 25 %, por lo que se decidió fijar el umbral en este valor. Una vez binarizado el índice manual, se aplicó umbrales (μ) a los índices automatizados, de forma que las ventanas con valores sobre el umbral fueran clasificadas como colaborativas.

De esta forma, se puede estudiar la relación entre ambas clasificaciones mediante su matriz de confusión y las métricas asociadas a esta. En este caso la calidad de la clasificación obtenida fue medida utilizando tres métricas comunmente usadas en clasificación binaria: *precisión, sensibilidad* y *exactitud*. La precisión y la sensibilidad miden la proporción de instancias correctamente clasificadas como positivas, con respecto a dos denominadores distintos. En el caso de la precisión, el denominador es la cantidad total de instancias clasificadas como positivas, mientras que en la sensibilidad se utiliza el total de instancias que eran realmente positivas. Por último, la exactitud mide la proporción entre las instancias correctamente clasificadas (positivas y negativas) y el total de instancias.

La tabla 2.2 presenta los resultados para los distintos umbrales utilizados para el número de hablantes activos, mientras que la tabla 2.3 presenta los resultados obtenidos utilizando el índice generado mediante PCA. En general, ambos índices presentaron máxima exactitud en un 67 %, valores aceptables en un problema con clases balanceadas. Además, como era esperable, en ambos casos se observaba un intercambio entre sensibilidad y presición al aumentar el umbral. Sin embargo, no se encontraron valores que permitieran optimizar ambas métricas al mismo tiempo.

Tabla 2.2: Precisión (P), sensibilidad (S) y exactitud (E) obtenidas al clasificar colaboración utilizando el número de hablantes activos para distintos valores del umbral μ. Mejores resultados en las métricas de evaluación en negrita.

μ	0.00	1.00	2.00	3.00	4.00
P	0.55	**0.64**	0.74	0.70	0.00
S	0.98	**0.91**	0.56	0.15	0.00
E	0.55	**0.67**	0.65	0.50	0.45

Tabla 2.3: Precisión (P), sensibilidad (S) y exactitud (E) obtenidas al clasificar colaboración utilizando el índice basasdo en PCA para distintos valores del umbral μ. Mejores resultados en las métricas de evaluación en negrita.

μ	-3.54	-2.85	-2.16	-1.47	-0.78	-0.09	0.60	1.29	1.98	2.67
P	0.55	0.55	0.55	0.55	0.57	**0.65**	0.72	0.67	0.71	0.00
S	1.00	0.99	0.98	0.98	0.96	**0.89**	0.66	0.25	0.05	0.00
E	0.55	0.55	0.56	0.56	0.59	**0.68**	0.67	0.52	0.47	0.45

Esto puede deberse a que los índices desarrollados pueden distinguir cuando los alumnos están interactuando entre sí, mas no pueden discernir si dicha interacción aporta a la resolución del problema o no, lo cual sí es capturado en le índice manual. A pesar de esto, los altos valores de sensibilidad para valores bajos del umbral en conjunto con valores aceptables de precisión en este mismo rango, sugieren que estos índices pueden ser valiosos como un primer filtro o una condición necesaria para la presencia de colaboración en una ventana. En

particular considerando que el umbral puede ser fijado de forma de establecer la proporción en la que se está dispuesto a intercambiar falsos positivos por falsos negativos.

2.3. Discusión

En este capítulo se han propuesto dos formulaciones para para obtener dos índices automatizados de colaboración a lo largo del tiempo utilizando diarización de voz y se estudió los efectos que distintos valores en los parámetros de estas formulaciones tienen sobre los índices obtenidos. Posteriormente, se estudiaron las curvas descritas por estos para distintas sesiones de RCP y las diferencias entre estas. Por último, se compararon los índices propuestos con respecto de la clasificación humana en un subconjunto de las sesiones utilizadas para desarrollarlos.

En particular, se encontró que mediante los índices automatizados fue posible idenfiticar dos grupos entre las sesiones en función de sus patrones en la toma de turnos. El primero mostraba una discusión inicial extendida del problema, seguido de poca interacción entre los alumnos con eventuales intervenciones en intervalos cortos de tiempo, lo que sugiere una forma de trabajo más individual en el grupo. Por otro lado, el segundo patrón mostraba una interacción constante a lo largo del tiempo, lo que puede implicar un enfoque más global a la hora de resolver el problema propuesto. De esta forma, los índices propuestos pueden ser una herramienta útil para investigadores que permita visualizar rápidamente la interacción del grupo a lo largo del tiempo, sin necesidad de estudiar detalladamente la sesión.

Adicionalmente, durante la comparación entre el índice manual y los índices automatizados, también se pudo observar similitud entre la cantidad de tiempo colaborativo en una ventana dada y los índices. Esto se puede observar particularmente en términos de los intevalos de crecimiento y decrecimiento de ambas curvas y en las posiciones de los mínimos y máximos locales. Más aún, tras plantear el problema como uno de clasificación, se encontró que ambos índices automatizados lograban valores de exactitud sobre el 65 % al clasificar las ventanas en colaborativas y no-colaborativas, resultados aceptables considerando la naturaleza no supervisada del problema.

A pesar de esto, las formulaciones propuestas presentaron algunas limitaciones. En ambos casos el intercambio entre precisión y sensibilidad al modificar el valor del umbral impedía optimizar ambas métricas al mismo tiempo. Adicionalmente, en ambos casos la máxima precisión alcanzada rondó el 70 % con valores muy bajos en la sensibilidad, mientras que la máxima sensibilidad estuvo sobre el 90 % mateniendo valores sobre el 50 % de precisión. Es decir, los índices automatizados en general reconocen la mayoría de las interacciones colaborativas, pero también son propensos a decir que hay interacciones productivas cuando no las hay.

Esto puede deberse a que ambos son buenos indicadores de la participación del grupo dada su naturaleza, sin embargo no permiten diferenciar entre interacciones que aporten al desarrollo de ideas sobre el problema e interacciones que no. Con esto en mente, se espera que una linea interesante para futura investigación en este campo pueda considerar el añadir información lingüística a la formulación de los índices automatizados, bajo la hipótesis de que dicha información ayudaría a identificar las interacciones que efectivamente se relacionan

con el problema propuesto.

En resumen, a pesar de no poder distinguir entre interacciones relativas al problema o no, los índices automatizados presentan dos ventajas importantes que los plantean como herramientas útiles para analizar sesiones de RCP. Primero, presentan una forma rápida y sencilla para visualizar las dinámicas de la interacción entre alumnos a lo largo de la sesión. Esto permite obtener una visión general de estas y estudiar patrones entre distintas sesiones sin necesidad de invertir tiempo en análisis detallados. Segundo, dados los altos valores de sensibilidad obtenidos, estos pueden ser usados para reducir rápidamente la cantidad de tiempo a estudiar mediante el descarte de ventanas con bajos valores en los índices. Por estas razones, se espera que el trabajo desarrollado pueda ser útil a académicos e investigadores interesados en el área de RCP.

Recaptulación y Conclusiones

En este trabajo de tésis se ha propuesto dos aplicaciones del aprendizaje automático, y particularmente del aprendizaje no-supervisado, al estudio de la RCP, las cuales se espera que en un futuro puedan convertirse en herramientas útiles para los investigadores de este área.

En primer lugar, se ha implementado un sistema que emplea modelos de ASR y NLP para identificar los tópicos subyacentes en un conjunto de sesiones y permite describir estas a lo largo del tiempo en función de los tópicos. Este sistema agrupa las frases en distintos tópicos basado en su similitud semántica y los dota de interpretabilidad encontrando sus palabras más representativas. Este proceso reveló tópicos relacionados contenidos específicos (Geometría, Números) y a formas específicas de interacción, como el compartir pantalla y el utilizar la pizarra. Similarmente, estos resultados fueron posteriormente reafirmados tras analizar las frases más representativas de cada uno.

Luego, las sesiones fueron descritas ordenando las frases de forma temporal, lo cual permitió estudiar la variación de los tópicos a lo largo de estas. Tras observar las visualizaciones de las sesiones, se encontró una preponderancia notoria de los tópicos de geometría y números en las sesiones de sus respectivos cursos. Este hallazgo motivó el estudiar diferencias entre los cursos para todos los tópicos. Por lo tanto, se realizó un test χ^2 de independencia sobre las proporciones en que los tópicos aparecían en cada curso, encontrando diferencias significativas no sólo en los tópicos de contenido, sino también en los tópicos de las distintas formas de interacción.

En segundo lugar, se propusieron dos índices automatizados de colaboración utilizando diarización de voz para obtener los turnos de cada estudiante y aplicando una batería de estadísticos sobre ventanas móviles. Posteriormente, se analizó las curvas obtenidas encontrando dos grupos mayoritarios de comportamiento. El primero presentaba una discusión inicial extendida, seguida de poca interacción con eventuales máximos locales, lo cual podría atribuirse a un trabajo más bien cooperativo. Por otro lado, el segundo patrón presentaba una interacción constante a lo largo del tiempo, que podría atribuirse a un enfoque más colaborativo, pero siendo más propenso a interacciones no productivas dada la cantidad de estas.

Tras analizar las curvas descritas por ambos índices, estas fueron comparadas con la anotación humana sobre un subconjunto de sesiones codificadas por un experto. Las visualizaciones generadas revelaron una similitud en términos de los mínimos y máximos locales entre los índices y las interacciones anotadas. Esto motivó a discretizar los índices utilizando un umbral y plantear el problema como uno de clasificación, considerando la naturaleza discreta de la anotación manual. Los resultados presentaron valores aceptables de exactitud y precisión

considerando la naturaleza no-supervisada de la metodología y buenos valores de sensibilidad, mostrando una dificultad para distinguir entre interacciones productivas y no-productivas, lo que concuerda con lo planteado en [43].

Los resultados y limitaciones encontrados motivan de forma natural dos posibles lineas futuras de investigación, las cuales se listan a continuación. En primer lugar, considerando otros resultados encontrados aplicando NLP en RCP [9],[10], [11]; es interesante estudiar una posible relación entre las representaciones obtenidas en el segundo capítulo y otras variables comunes de interés, como podrían ser las ganancias de aprendizaje de los estudiantes o el desempeño final del grupo en el problema. En segundo lugar, dada la dificultad que los índices de colaboración planteados en el tercer capítulo presentan para distinguir las interacciones productivas, es posible que el acoplar una representación más basada en el contenido lingüístico al cálculo de estos permita mejorar su desempeño en la tarea de clasificación.

En síntesis, esta tesis presenta nuevas aplicaciones del aprendizaje no-supervisado al estudio de la RCP, implementa dos representaciones automatizadas para facilitar el análisis de las sesiones de aprendizaje colaborativo, y presenta resultados interesantes sobre el potencial que esta tecnología puede entregar en este área. Adicionalmente, las representaciones desarrolladas presentan un mecanismo para analizar de forma más general grandes cantidades de sesiones en poco tiempo, sin necesidad de un equipo de codificadores y con requerimientos básicos de audio. Por estas razones, se espera que en un futuro estas puedan convertirse en una herramienta útil y valiosa para investigadores de este área.

Bibliografía

[1] OECD, PISA 2015 Assessment and Analytical Framework. 2017, doi:https://doi.org/https://doi.org/10.1787/9789264281820-en.

[2] Fiore, S. M., Graesser, A., y Greiff, S., "Collaborative problem-solving education for the twenty-first-century workforce," Nature human behaviour, vol. 2, no. 6, pp. 367–369, 2018.

[3] Peña-López, I. *et al.*, "Pisa 2015 results (volume v). collaborative problem solving," 2017.

[4] Griffin, P. y Care, E., "The atc21s method," en Assessment and teaching of 21st Century Skills, pp. 3–33, Springer, 2015.

[5] Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., y Hesse, F. W., "Advancing the science of collaborative problem solving," Psychological Science in the Public Interest, vol. 19, no. 2, pp. 59–92, 2018.

[6] Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., Massey, C., O'Neil, H., Pellegrino, J., Rothman, R., *et al.*, "Collaborative problem solving: Considerations for the national assessment of educational progress," 2017.

[7] Kang, J., An, D., Yan, L., y Liu, M., "Collaborative problem-solving process in a science serious game: Exploring group action similarity trajectory.," International Educational Data Mining Society, 2019.

[8] Reilly, J. M. y Schneider, B., "Predicting the quality of collaborative problem solving through linguistic analysis of discourse.," International Educational Data Mining Society, 2019.

[9] Pugh, S. L., Subburaj, S. K., Rao, A. R., Stewart, A. E., Andrews-Todd, J., y D'Mello, S. K., "Say what? automatic modeling of collaborative problem solving skills from student speech in the wild.," International Educational Data Mining Society, 2021.

[10] Stewart, A. E., Vrzakova, H., Sun, C., Yonehiro, J., Stone, C. A., Duran, N. D., Shute, V., y D'Mello, S. K., "I say, you say, we say: Using spoken language to model socio-cognitive processes during computer-supported collaborative problem solving," Proceedings of the ACM on Human-Computer Interaction, vol. 3, no. CSCW, pp. 1–19, 2019.

[11] Stewart, A. E., Keirn, Z., y D'Mello, S. K., "Multimodal modeling of collaborative problem-solving facets in triads," User Modeling and User-Adapted Interaction, vol. 31, no. 4, pp. 713–751, 2021.

[12] Lai, E., DiCerbo, K., y Foltz, P., "Skills for today: What we know about teaching and assessing collaboration.," Pearson, 2017.

[13] Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., y Fischer, F., "Analyzing collaborative learning processes automatically: Exploiting the advances of

computational linguistics in computer-supported collaborative learning," International journal of computer-supported collaborative learning, vol. 3, no. 3, pp. 237–271, 2008.

[14] Hao, J., Chen, L., Flor, M., Liu, L., y von Davier, A. A., "Cps-rater: Automated sequential annotation for conversations in collaborative problem-solving activities," ETS Research Report Series, vol. 2017, no. 1, pp. 1–9, 2017.

[15] Oviatt, S. y Cohen, A., "Written and multimodal representations as predictors of expertise and problem-solving success in mathematics," en Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 599–606, 2013.

[16] Murray, G. y Oertel, C., "Predicting group performance in task-based interaction," en Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 14–20, 2018.

[17] Chopade, P., Edwards, D., Khan, S. M., Andrade, A., y Pu, S., "Cpsx: Using ai-machine learning for mapping human-human interaction and measurement of cps teamwork skills," en 2019 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–6, IEEE, 2019.

[18] Samrose, S., Zhao, R., White, J., Li, V., Nova, L., Lu, Y., Ali, M. R., y Hoque, M. E., "Coco: Collaboration coach for understanding team dynamics during video conferencing," Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, vol. 1, no. 4, pp. 1–24, 2018.

[19] Dowell, N. M., Nixon, T. M., y Graesser, A. C., "Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions," Behavior research methods, vol. 51, no. 3, pp. 1007–1041, 2019.

[20] Dowell, N. M., Lin, Y., Godfrey, A., y Brooks, C., "Exploring the relationship between emergent sociocognitive roles, collaborative problem-solving skills, and outcomes: A group communication analysis.," Journal of Learning Analytics, vol. 7, no. 1, pp. 38–57, 2020.

[21] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., y Gill, M.-P., "pyannote.audio: neural building blocks for speaker diarization," en ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2020.

[22] Li, M., Zhou, S., y Xu, B., "Long-running speech recognizer: An end-to-end multi-task learning framework for online asr and vad," arXiv preprint arXiv:2103.01661, 2021.

[23] Schlotterbeck, D., Jiménez, A., Araya, R., Caballero, D., Uribe, P., y Van der Molen Moris, J., ""teacher, can you say it again?ïmproving automatic speech recognition performance over classroom environments with limited data," en International Conference on Artificial Intelligence in Education, pp. 269–280, Springer, 2022.

[24] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., y Weber, G., "Common voice: A massively-multilingual speech corpus," arXiv preprint arXiv:1912.06670, 2019.

[25] Grosman, J., "Fine-tuned XLSR-53 large model for speech recognition in Spanish." https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-spanish, 2021.

[26] AI, M., "Papers with code: A free and open resource with machine learning papers, code, datasets, methods and evaluation tables.," 2022, https://paperswithcode.com/sota/sp

eech-recognition-on-common-voice-spanish (visitado el 2022-07-14).

[27] Panayotov, V., Chen, G., Povey, D., y Khudanpur, S., "Librispeech: an asr corpus based on public domain audio books," en 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5206–5210, IEEE, 2015.

[28] Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., y Dupoux, E., "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," arXiv preprint arXiv:2101.00390, 2021.

[29] Kohavi, R. *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," en Ijcai, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

[30] Hawkins, D. M., "The problem of overfitting," Journal of chemical information and computer sciences, vol. 44, no. 1, pp. 1–12, 2004.

[31] McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., y Bourlard, H., "On the use of information retrieval measures for speech recognition evaluation," rep. tec., IDIAP, 2004.

[32] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., *et al.*, "Transformers: State-of-the-art natural language processing," en Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp. 38–45, 2020.

[33] Heafield, K., "Kenlm: Faster and smaller language model queries," en Proceedings of the sixth workshop on statistical machine translation, pp. 187–197, 2011.

[34] LLC, K. T., "pyctcdecode," 2022, https://github.com/kensho-technologies/pyctcdecode (visitado el 2022-07-14).

[35] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., y Bowman, S. R., "Glue: A multi-task benchmark and analysis platform for natural language understanding," arXiv preprint arXiv:1804.07461, 2018.

[36] Pérez, J. M., Giudici, J. C., y Luque, F., "pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks," 2021.

[37] Bellman, R., "Dynamic programming," Science, vol. 153, no. 3731, pp. 34–37, 1966.

[38] Taylor, C. R., "Dynamic programming and the curses of dimensionality," en Applications of dynamic programming to agricultural decision problems, pp. 1–10, CRC Press, 2019.

[39] McInnes, L., Healy, J., y Melville, J., "Umap: Uniform manifold approximation and projection for dimension reduction," arXiv preprint arXiv:1802.03426, 2018.

[40] Schwarz, G., "Estimating the dimension of a model," The annals of statistics, pp. 461–464, 1978.

[41] Bholowalia, P. y Kumar, A., "Ebk-means: A clustering technique based on elbow method and k-means in wsn," International Journal of Computer Applications, vol. 105, no. 9, 2014.

[42] Özgür, A., Özgür, L., y Güngör, T., "Text categorization with class-based and corpus-based keyword selection," en International Symposium on Computer and Information Sciences, pp. 606–615, Springer, 2005.

[43] Hu, L. y Chen, G., "Exploring turn-taking patterns during dialogic collaborative problem solving," Instructional Science, vol. 50, no. 1, pp. 63–88, 2022.

[44] Lemke, J. L., Talking science: Language, learning, and values. ERIC, 1990.

[45] Gibson, D. R., "Taking turns and talking ties: Networks and conversational interaction," American journal of sociology, vol. 110, no. 6, pp. 1561–1597, 2005.

[46] Hu, L., "Turn-usurping in dialogic collaborative problem solving," en Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021., International Society of the Learning Sciences, 2021.

[47] Choi, H. y Kang, M., "Applying an activity system to online collaborative group work analysis," British Journal of Educational Technology, vol. 41, no. 5, pp. 776–795, 2010.

[48] Heo, H., Lim, K. Y., y Kim, Y., "Exploratory study on the patterns of online interaction and knowledge co-construction in project-based learning," Computers & Education, vol. 55, no. 3, pp. 1383–1392, 2010.

[49] Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., y Narayanan, S., "A review of speaker diarization: Recent advances with deep learning," Computer Speech & Language, vol. 72, p. 101317, 2022.

[50] Yin, R., Bredin, H., y Barras, C., "Neural speech turn segmentation and affinity propagation for speaker diarization," en Annual Conference of the International Speech Communication Association, 2018.

[51] De Boer, P.-T., Kroese, D. P., Mannor, S., y Rubinstein, R. Y., "A tutorial on the cross-entropy method," Annals of operations research, vol. 134, no. 1, pp. 19–67, 2005.

[52] Bredin, H., "Tristounet: triplet loss for speaker turn embedding," en 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5430–5434, IEEE, 2017.

[53] Hoffer, E. y Ailon, N., "Deep metric learning using triplet network," en International workshop on similarity-based pattern recognition, pp. 84–92, Springer, 2015.

[54] Schroff, F., Kalenichenko, D., y Philbin, J., "Facenet: A unified embedding for face recognition and clustering," en Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823, 2015.

[55] Madhulatha, T. S., "An overview on clustering methods," arXiv preprint arXiv:1205.1117, 2012.

[56] Dueck, D., Affinity propagation: clustering data by passing messages. Citeseer, 2009.

[57] Abdi, H. y Williams, L. J., "Principal component analysis," Wiley interdisciplinary reviews: computational statistics, vol. 2, no. 4, pp. 433–459, 2010.