Palabras y Silencios es la Edición Digital de la Asociación Internacional de Historia Oral. Incluye artículos de un rango variado de disciplinas y es una medio para que la comunidad profesional comparta proyectos y tendencias actuales en la historia oral alrededor del mundo

http://ioha.org
Online ISSN 2222-4181

Palabras y Silencios
Septiembre de 2018
"Memoria y narración"

Audio Mining:
Advanced Speech Analytics for Oral History[1]

Almut Leh, Michael Gref and Joachim Köhler
Fernuniversität in Hagen
almut.leh@fernuni-hagen.de

1. Introducción: Qué pueden hacer las tecnologías para la investigación en humanidades

Los datos audiovisuales (multimodales) juegan un papel crecientemente importante en humanidades y ciencias sociales. De hecho, todas las humanidades y ciencias sociales entran en contacto ocasionalmente con estos datos. El enorme progreso hecho en el desarrollo de aparatos digitales de grabación, que todavía hoy permite realizar grabaciones audiovisuales con menor gasto personal y financiero comparativamente hablando, sugiere que estos datos van a ser cada vez más importantes a corto y medio plazo. Comparadas con datos textuales, las posibilidades de la evaluación, documentación, publicación (generalmente disposición de accesibilidad) y archivo de estos datos audiovisuales multimodales que utilizan las modernas tecnologías de la información están todavía poco estudiados e investigados.[2] En el presente, la diseminación científica y el uso de estos datos continúan basándose fundamentalmente en transcripciones, p. ej, convirtiendo en el medio de lenguaje escrito. Esto tiene dos desventajas cruciales: Primero, la producción de transcripciones es un trabajo intensivo y que consume mucho tiempo. Segundo, las transcripciones pierden las características específicas de la melodía del discurso, cualidades vocales, gestos y expresiones faciales, etc, que son específicas de los datos audiovisuales. Sin embargo, estas últimas son a menudo decisivas

[1] El texto presenta los resultados de un Proyecto de investigación en marcha que tiene como objetivo desarrollar tecnologías del lenguaje para metodologías basadas en entrevistas. Los escenarios de ejemplo son el campo de la entrevista linguistica y la entrevista de historia oral. El proyecto "KA³" está financiado por el Ministerio Federal de Educación e Investigación. El grupo del proyecto está formado por Nikolaus P. Himmelmann (Instituto de Lingüistica, Universidad de Colonia) a la coordinación, Joachim Köhler y Michael Gref (Instituto Fraunhofer para Análisis de Inteligencia y Sistemas de Información (IAIS) y Almut Leh (FernUniversität de Hagen). para más información, ver Leh, Almut, Joachim Köhler y Nikolaus P. Himmelmann: Speech analytics in research based on qualitative interviews. Experiences from KA3. *VIEW* Special Issue *Audiovisual Data in Digital Humanities* (2019).

[2] Comparar, por ejemplo, el siguiente libro de texto: Fotis Jannidis, Hubertus Kohle, Malte Rehbein (eds.): Digital Humanities. Eine Einführung, Stuttgart 2017.

para la interpretación apropiada de lo que se ha dicho. Aquí es donde entra el audio minado (audiomining).

Al utilizar tecnología de minado de audio el manejo de los datos audiovisuales debería ser más eficiente, especialmente en dos maneras: (a) reduciendo recursos típicamente empleados para el análisis preparatorio de entrevistas y (b) minimizando las reducciones inherentes a la transcripción. Además, esperamos que el minado de audio no sólo facilite el proceso de investigación, sino que también permita nuevos acercamientos en lo que respecta a cantidad y calidad. Los estudios comparativos y los análisis cuantitativos son más razonables al cubrir más datos. Estas tecnologías de discurso también representan aspectos verbales y no verbales de la comunicación de modo más profundo y sistemático, abriendo así nuevas dimensiones para el análisis cualitativo.

2. La necesidad de tecnología de discursos en el dominio de la historia oral

En Alemania, así como en otros países europeos, el uso de historia oral como metodología desde los años 70 ha resultado en una contribución sustancial de entrevistas y proyectos de historia oral al registro histórico. El legado material de todos estos proyectos desemboca en miles de grabaciones de audio, así como una inmensa cantidad de grabaciones de video de distintos tipos de soporte de datos, que nos cuentan la historia técnica de décadas pasadas. Es característico para entrevistas de historia oral que puedan ser utilizadas para distintas preguntas e interpretaciones. Debido a su forma narrativa y su dimensión biográfica, pueden contener tanta abundancia de información que no es agotada en una sólo evaluación. Por esta razón las entrevistas de historia oral son frecuentemente sujeto de re análisis. Sobre todo, cuando el/la testigo contemporáneo/a ya ha fallecido, por lo que las entrevistas son el único modo de acceder a sus experiencias. El archivo y análisis de las entrevistas de historia oral realizadas por otras personas, sin embargo, presenta muchos desafíos. La recuperación de

entrevistas a partir de ciertas preguntas de investigación es concretamente un problema. La información biográfica puede ser obtenida con metadatos.

Sin embargo, tan pronto como el interés se dirige hacia ciertos contenidos, la recuperación es muy difícil y sólo posible en el mejor de los casos por medio de una búsqueda del texto completo.

La transcripción ha jugado un rol central tradicionalmente en el re análisis de las entrevistas. Es cierto que el análisis no debe limitarse a la transcripción, sino también incluir el modo de hablar, las expresiones faciales y los gestos. Pero también es cierto que la transcripción es una herramienta indispensable para el análisis y archivo de las entrevistas de historia oral. En otras palabras, la falta de transcripciones limitará ostensiblemente el uso de las entrevistas. En ese contexto, no sólo es un problema que la mayor parte de las entrevistas en los archivos no están transcritas,[3] sino que muchas de las entrevistas que hoy se realizan no tienen prevista una transcripción. Todas estas entrevistas son de uso limitado y serán olvidadas en muchos casos. Contra esta tendencia, es obvio que la investigación en historia oral de hoy y el futuro tiene una gran necesidad de tecnología para el reconocimiento automático de discursos y/o minado de audio.

3. El sistema de Audio Minado Fraunhofer IAIS

Mientras que la presencia cotidiana de siri, Alexa y otras "asistentes virtuales" similares, las tecnologías automáticas dan la impresión de que poderososos programas para el reconocimiento de voz se han convertido desde hace tiempo en el último grito, la experiencia en la práctica investigadora es claramente diferente. Las entrevistas de discurso espontáneo combinadas con técnicas de grabación no profesionales todavía suponen un gran desafío para

[3]En el archivo "Deutsches Gedächtnis" de la FernUniversität de Hagen, el mayor archive de entrevistas de Alemania, alrededor del 30% está sin transcriber. Ver: Almut Leh: Das Archiv „Deutsches Gedächtnis" und seine Bestände: Herkunft – Erschließung – Nutzung, in: Denis Huschka, Hubert Knoblauch, Claudia Oellers und Heike Solga (Hg.): Forschungsinfrastrukturen für die qualitative Sozialforschung. Berlin: SciVero, 2013, pp. 109-116.

las tecnologías del lenguaje. Mientras que el reconocimiento del discurso funciona muy bien en distintas áreas de la vida cotidiana, los resultados con entrevistas de historia oral son muy débiles.

Durante tres años hemos estado trabajando en los archivos “Deutsches Gedächtnis” de la Fernuniversität de Hagen, el mayor archivo de historia oral de Alemania, junto al Instituto Fraunhofer Institute para el Análisis Inteligente y los sistemas de información IAIS para mejorar el rendimiento de los programas de reconocimiento de discurso en entrevistas de historia oral. El Instituto Fraunhofer (IAIS) tiene una extensiva práctica en el campo del reconocimiento de voz, y lo ha desarrollado en una aplicación para archivos en el dominio de la banda ancha. El trabajo con entrevistas de Historia Oral es un nuevo reto para el IAIS, que conrtibuye a la mejora del reconocimiento del discurso como un conjunto. El objetivo del proyecto KA³es mejorar el reconocimiento del discurso hasta el punto de que pueda ser utilizado en la práctica y suponer un valor añadido. En los siguientes párrafos me gustaría presentar algunas herramientas de Audio Minado desarrolladas por el instituto y su puesta en práctica en la actualidad.

El sistema de Audio Minado Fraunhofer está diseñado para analizar automáticamente la señal audio de archivos de medios audiovisuales. El objetivo es crear una transcripción alineada en el tiempo de las palabras escritas, como lo hacen normalmente los profesionales humanos que se dedican a transcribir. Ello no sólo incluye el reconocimiento automático del discurso sino todo un flujo de trabajo por el que los archivos se estructuran y se hacen susceptibles de búsquedas de texto. En un nivel superior del sistema, también se puede generar un archivo adicional XML para cada audio, que contiene distintos tipos de códigos relativos a información.

1. Segmentación de Audio

2. Detección de conceptos

3. Agrupamiento de hablantes

4. Reconocimiento automático de discurso

5. Extracción de palabras clave

Hay varias etapas en este proceso. Primero, el discurso es automáticamente dividido en segmentos homogéneos. Después de la segmentación, cada segmento es clasificando utilizando un detector de discurso/no discurso. Los segmentos que se asume que contienen discursos se clasifican adicionalmente mediante un detector basado en la asignación de género a voces masculinas o femeninas. En el siguiente paso del proceso se aplica un agrupamiento. Aquí todos los segmentos de discurso del mismo hablante reciben el mismo número identificador. Todo el proceso de segmentación de discurso se resumen como proceso de diarización del hablante. El código temporal de cada segmento y ética se almacena el archivo de metadatos del MPEG-7 metadata file.

En el siguiente paso se aplica tecnología para el reconocimiento automático del discurso. Se ha hecho un progreso significativo en el reconocimiento del discurso que ha estado basado en la utilización de redes neuronales profundas. Estas redes han contribuido decisivamente al desarrollo de la inteligencia artificial en años recientes.[4] Algunos de los términos utilizados en

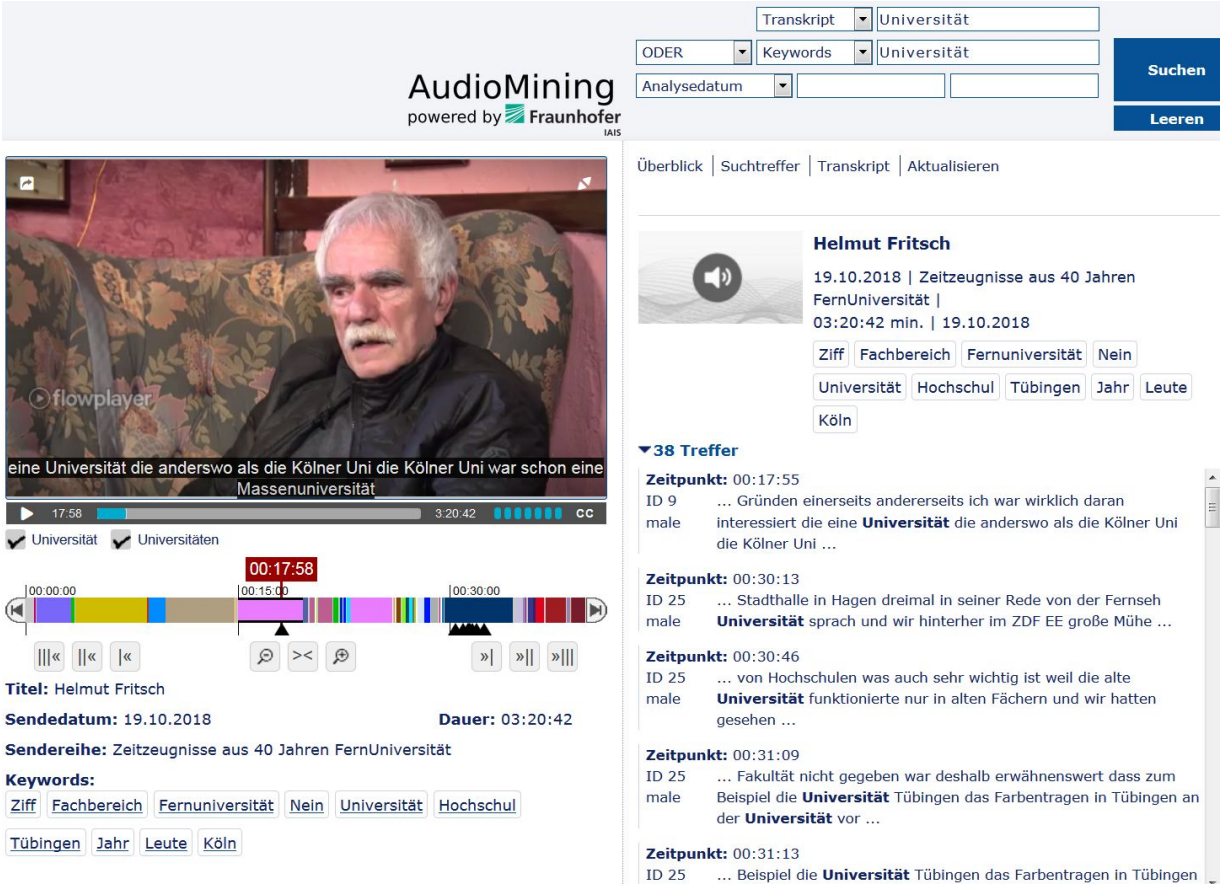
[4] Información especial relativa a los aspectos técnicos puede consultarse en Michael Gref, Joachim Köhler, Almut Leh: Improved Transcription and Indexing of Oral History Interviews for Digital Humanities Research, in: Nicoletta Calzolari, et al. ed.: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, May 2018, 3124-3131.
http://www.lrec-conf.org/proceedings/lrec2018/summaries/137.html

estos años son “aprendizaje profundo (“deep learning)” o “Aprendizaje automático” (“machine learning”).

El modelado acústico está extraído de 1000 horas de emisión grabadas en Alemania. El sistema de reconocimiento de discurso contiene un *léxico de pronunciación* para 500.000 palabras aproximadamente. Las transcripciones fonéticas de cada palabra se generan automáticamente utilizando un convertidor estadístico de grafemas a fonemas.

Las transcripciones finales generadas automáticamente proporcionan una mayor funcionalidad de búsqueda y navegación. Para cada palabra reconocida se genera un código en el tiempo exacto que es almacenado en el archivo de metadatos MPEG-7. La imagen muestra la interfaz de búsqueda y recuperación de datos del sistema de Minado de Audio.

Proporcionando una interfaz gráfica de usuario (Graphical User Interface-GUI) el sistema permite al usuario final navegar entre las entrevistas. La interfaz se muestra en esta imagen:



Graphical Web User Interface of the Fraunhofer IAIS Audio Mining system

Un reproductor de media incrustado permite al usuario reproducir directamente segmentos de hablantes concretos, representados por un color único para cada uno/a en la barra temporal debajo del vídeo. El usuario o usuaria puede buscar por ítems relevantes y adquirir directamente los cortes de las frases reconocidas y los puntos de entrada (triángulos negros) para saltar a la posición donde el ítem es mencionado. La aplicación proporciona además una lista de palabras clave relevantes. El resultado de las analíticas del proceso es desplegado en la secuencia de segmentación coloreada. La herramienta es capaz de procesar grabaciones de audio extremadamente largas, con una duración aproximada de cuatro horas.

4. Las entrevistas de historia oral como un desafío

Aplicados a escenarios de emisión de discurso, los acercamientos más avanzados al reconocimiento de voz logran un margen de error en palabras del 8%, lo que permite una comprensión significativa del texto. Esto permite una transcripciones de alta calidad que pueden ser utilizadas directamente con propósitos de lectura y comprensión de textos. Aplicando esta tecnología a la historia oral, los márgenes de error son mucho más elevados por los distintos niveles de habla y las dificultades relativas a las condiciones de la grabación. Las primeras pruebas lanzadas al inicio de nuestro proyecto resultaban en un margen de error del 60%. Pero tras cuarto años de investigación (mayo 2019) hemos sido capaces de reducir el margen de error al 25%. En vista de los distintos estilos y calidades de las grabaciones, no es sorprendente que los resultados varíen ampliamente. Los márgenes de error no superan el 15% en algunas entrevistas, mientras que en otras todavía son muy altos.

Uno/a podría esperar que la pronunciación y el dialecto utilizados por las partes de la entrevista sería el mayor problema. Lo cierto es que la calidad de la grabación es la que causa los mayores problemas. La señal de grabación de audio es uno de los mayores desafíos para adaptar el sistema ASR a las entrevistas de historia oral. Las señales de audio utilizadas para entrenar el modelo acústico fueron grabadas y procesadas con posterioridad utilizando un equipo profesional de alta calidad. Las grabaciones no tenían normalmente ruido de fondo, un eco apenas perceptible y niveles bien ajustados. Las entrevistas de historia oral trabajadas habían sido realizadas mayoritariamente “en el campo”, por ejemplo, el salón de una entrevistada, la cocina, u otros espacios donde los ruidos externos son menos controlables. El uso de micrófonos de mesa contamina las grabaciones con eco, lo que no es un problema para el oído humano pero sí que interfiere con los programas de reconocimiento de voz.

Hay dos vías para hacer frente a estos desafíos: Mejora del discurso y Entrenamiento Multi-Condicional. La primera apunta a modificar la señal en sí misma e incrementar la calidad del audio al reducir ruidos y compensar corrupciones. El entrenamiento Multi-Condicional, por otra parte, apunta a robustecer el modelo acústico frente a las corrupciones mostrando al modelo un amplio rango de características de señales de audio corruptas durante el entrenamiento. Este acercamiento apunta a dejar que el modelo generalice a distorsiones y tipos de ruido imperceptibles durante el entrenamiento. Ello es posible al confiar solamente en características acústicas solidas. Fraunhofer utiliza el último enfoque para tratar de generar un modelo acústico robusto.

Pero por supuesto, otros grandes retos planteados por la historia oral tienen que ver con el uso de lenguaje coloquial en discurso espontáneo, dudas, cambios en el modo de hablar vinculados a la edad o la salud, palabras específicas utilizadas durante la entrevista que raramente se emplean en la vida cotidiana, (p. ej *Kriegerwitwenso¨hne*, palabra alemana para referirse a *los hijos de una viuda de guerra*). Estos desafíos deben ser encarados adaptando el modelo del lenguaje y serán trabajo a realizar en el futuro.

5. Conclusión y Perspectivas

Aunque los resultados del reconocimiento de discurso no son suficientes en la actualidad para la producción automática de transcripciones incluso con buen material original, el reconocimiento de discurso puede utilizarse como una ventaja. Por un lado, las transcripciones incorrectas generadas por esta herramienta pueden corregirse manualmente, salvando tiempo y recursos en comparación con una transcripción íntegra. Por otro lado, estas herramientas permiten una mejora de la práctica de archivo. El acceso directo a la señal de audio y la pre-estructuración del contenido permiten la recuperación de grandes cantidades de información, por lo que los datos específicos pueden hacerse disponibles para requerimientos

de la investigación en un sentido amplio. En este sentido, el reconocimiento de discurso nos permite incluir entrevistas no transcritas en las búsquedas, por lo que pueden utilizarse para análisis secundarios, mientras que de lo contrario estarían perdidas para futuras investigaciones. Al disponer los resultados de la búsqueda en un contexto, su relevancia es incrementada al momento, incrementando la eficiencia de la búsqueda. Las palabras clave generadas automáticamente también son un gran avance para la práctica del archivo cuando se trata de encontrar entrevistas o fragmentos relevantes al margen del estatus de la transcripción. A la vista de los márgenes de error actuales en el reconocimiento de discurso, la capacidad de buscar por términos y palabras clave está incompleta, pero ya han añadido un valor de modo significativo, porque han permitido incluir un gran número de entrevistar sin transcribir en las búsquedas.

También en el análisis, el acceso directo a la señal es una ganancia considerable para la historia oral y la investigación biográfica, p. ej la implementación de la reclamación anteriormente satisfecha de tratar el audio como una fuente primaria e incluir entonces el modo de hablar (tono de voz, calidad de la voz, pausas, etc) en la interpretación. De hecho, el análisis de las entrevistas se basa a menudo solamente en el conocimiento de la transcripción, lo que puede ser causa de malinterpretaciones, especialmente en la evaluación de entrevistas realizadas por otros. El acceso a la señal de audio, por otro lado, permite una representación sincrónica del audio u la transcripción, por lo que la recepción de la fuente primaria es un requisito previo a la interpretación de la entrevista y las intenciones manifestadas.

Las herramientas, adems, abrieron nuevas dimensiones para preguntas de investigación. Mientras que el número de casos para el análisis convencional es de unas treinta entrevistas, números mucho mayores pueden procesarse con el apoyo técnico y evaluar entonces cuantitativamente una gran cantidad de preguntas comparativas. Al mismo tiempo, las

herramientas también ofrecen nuevas dimensiones para el análisis cualitativo, en el que los aspectos lingüisticos y no linguisticos de la comunicación puede grabarse, documentarse y ser accesibles a preguntas de investigación de manera diferenciada. Estudios piloto realizados en el futuro muestran que posibilidades concretas ofrecen estos acercamientos cualitativos y cuantitativos. El uso de tecnologías de reconocimiento de voz/discurso en investigaciones basadas en entrevistas se ha convertido en algo más que una promesa. El reconocimiento automático ya está siendo utilizado con éxito en el análisis y los archivos. Por encima de todo, no obstante, quedan claras las ventajas que ofrecen las tecnologías del lenguaje. En la actualidad existen investigaciones dedicadas a averiguar cómo modelar las condiciones cambiantes y desafiantes de las entrevistas de historia oral. Se están aplicando entrenamientos multicondicionales dirigidos a robustecer los modelos acústicos. La conclusión que nos gustaría plantear es que se ha logrado mucho y que mucho más necesita hacerse todavía.