

Inspirados por la naturaleza multisensorial de la percepción del habla humana, ha habido un creciente interés en el desarrollo de sistemas automáticos de lectura de labios. Esta desafiante tarea tiene como objetivo interpretar el habla analizando únicamente los movimientos labiales del hablante sin depender de señales acústicas. Entre la amplia variedad de aplicaciones ofrecidas por esta tecnología, destacamos su prometedor papel en el avance de las interfaces de habla silenciosas a través del diseño de soluciones no invasivas que den soporte a la comunicación de personas que han perdido su capacidad de hablar.

La eficacia de esta tecnología, sin embargo, se ve comprometida por numerosos factores, entre los que se incluyen las ambigüedades visuales a la hora de intentar distinguir entre sonidos distintos pero visualmente similares, y la dificultad a la que se enfrentan los modelos multi-hablantes para generalizar y abordar las variaciones inter-personales entre hablantes. Aparte de estos desafíos inherentes a la lectura de labios en sí misma, existen cuestiones más técnicas, como son la escasez de datos audiovisuales para el habla continua y natural en idiomas distintos del inglés y la actual dependencia de modelos de alta complejidad para lograr un rendimiento sólido. En relación con el desarrollo de interfaces de habla silenciosa, también observamos que la investigación sobre enfoques dependientes del hablante se ha limitado en su mayoría a entornos controlados y protocolos experimentales restringidos, pasando por alto la importancia de explorar este tipo de estudios dependientes del hablante en condiciones más naturalistas, en particular con habla continua.

Esta tesis no sólo aborda esta brecha de investigación en la dependencia del hablante, sino que también investiga el diseño de sistemas automáticos de lectura labial y su integración con señales acústicas del habla para el español continuo natural. Nuestro enfoque integral ofrece una perspectiva holística del problema, comprendiendo la recopilación y el procesamiento de datos audiovisuales, la transición de los métodos tradicionales basados en modelos ocultos de Markov a arquitecturas profundas de extremo a extremo de última generación, y la integración de señales del habla audiovisuales para el reconocimiento robusto de la voz. Más específicamente, proponemos estrategias de entrenamiento inter-modal en el contexto de decodificadores tradicionales, evaluamos y analizamos la eficacia de las arquitecturas de extremo a extremo actuales y su adaptación a hablantes específicos, realizamos estudios comparativos sistemáticos entre estos dos distintos paradigmas de decodificación en diversas condiciones de entrenamiento, y exploramos las contribuciones de las señales visuales cuando se integran con señales acústicas para el reconocimiento de voz audiovisual bajo una perspectiva de eficiencia. Como resultado, nuestra investigación establece un punto de referencia de lectura labial en español, con el objetivo de garantizar que las tecnologías de lectura labial se puedan estudiar de manera efectiva para este idioma en diversas condiciones de datos.

Un factor común en todas estas contribuciones es nuestro particular énfasis en la eficiencia de esta tecnología y su aplicación en situaciones del mundo real. Al abordar la heterogeneidad y la falta ocasional de datos que sufren la mayoría de los idiomas en el contexto de las tecnologías de habla audiovisual, creemos firmemente que nuestro trabajo promueve una mayor investigación en idiomas subrepresentados y entornos de bajos recursos.