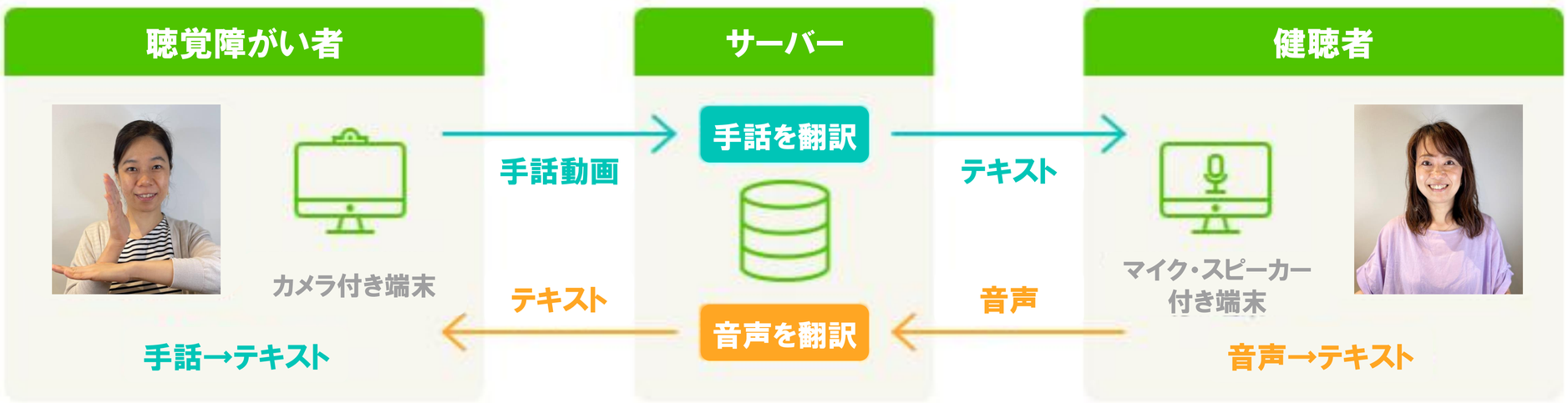
手話と音声の双方向コミュニケーションシステム（SureTalk）
～メディア発信とe-learning開発に向けて～

［共同研究開発］
国立大学法人電気通信大学 ｜ ソフトバンク株式会社

SureTalkのシステム構成



SureTalkの仕組みを実現するために、最先端のAI技術を活用しています

手話動画を認識する技術

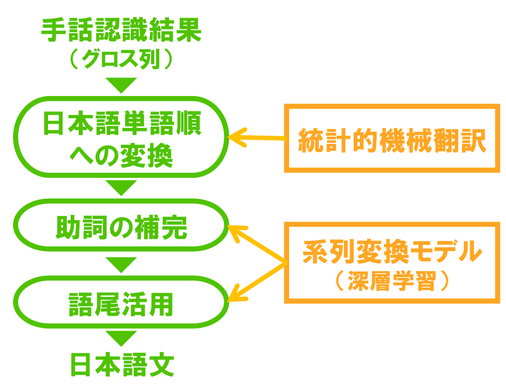
手話認識部では、パソコンやスマートフォンなどのカメラを用いて撮影した動画像に対して、深層学習を活用した身体動作の追跡処理を行い、話者の骨格座標を得ます。その後、骨格座標に基づいて話者の姿勢や動作など手話を認識するために重要な特徴を抽出します。最後に、深層学習を活用した時系列認識処理によって話者が表現している手話の単語順を認識します。深層学習による身体動作追跡を用いることで、服装をモノトーンにすることや撮影の背面をグリーンにすること、さらに深度センサーを利用することなどの特殊な条件が不要となり、日常的な環境での撮影が可能になりました。さらに、時系列認識処理に深層学習を適用することで、手話の認識精度を向上させています。

（担当：電気通信大学情報理工学研究科情報学専攻 高橋 裕樹研究室）

単語順を対象に自然言語処理を行い文章化する技術

自然言語処理部では、手話単語順と日本語文の対訳例から手話単語と日本語単語との対応関係、助詞の補完、および動詞・助動詞の活用について統計的機械翻訳の手法や深層学習を用いた翻訳モデルを学習します。その翻訳モデルを使用して、手話認識部から出力される手話単語順を変換後の日本語単語順の生起確率が最大になるように翻訳を行います。これにより、手話単語順をより自然な日本語文に変換することが可能になりました。

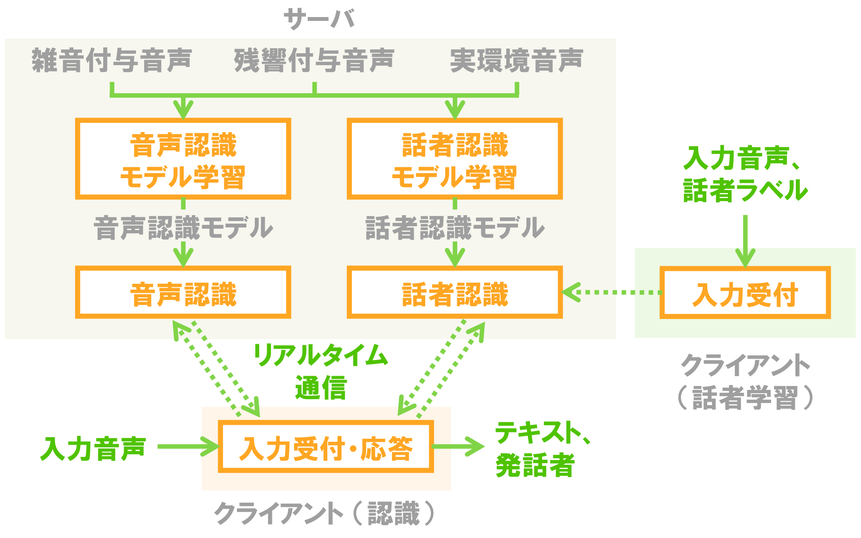
（担当：電気通信大学情報理工学研究科情報学専攻 内海 彰研究室）



音声認識に関わる開発

音声処理部では、オープンソースの深層学習に基づく音声認識エンジンを利用して、健聴者が発声した音声情報をリアルタイムに文字情報に変換します。雑音のある環境、反響の多い環境などのさまざまな環境における音声データを独自に収集し、それらを用いて音響モデルを学習させることで、実環境下における認識精度を高めることが可能になりました。また、話者識別機能を実装し、誰が発声したかを判別できるようにすることで、複数話者におけるコミュニケーションを円滑化することを目指しています。

（担当：電気通信大学情報理工学研究科情報・ネットワーク工学専攻 中鹿 亘研究室）



国立大学法人 電気通信大学 The University of Electro-Communications

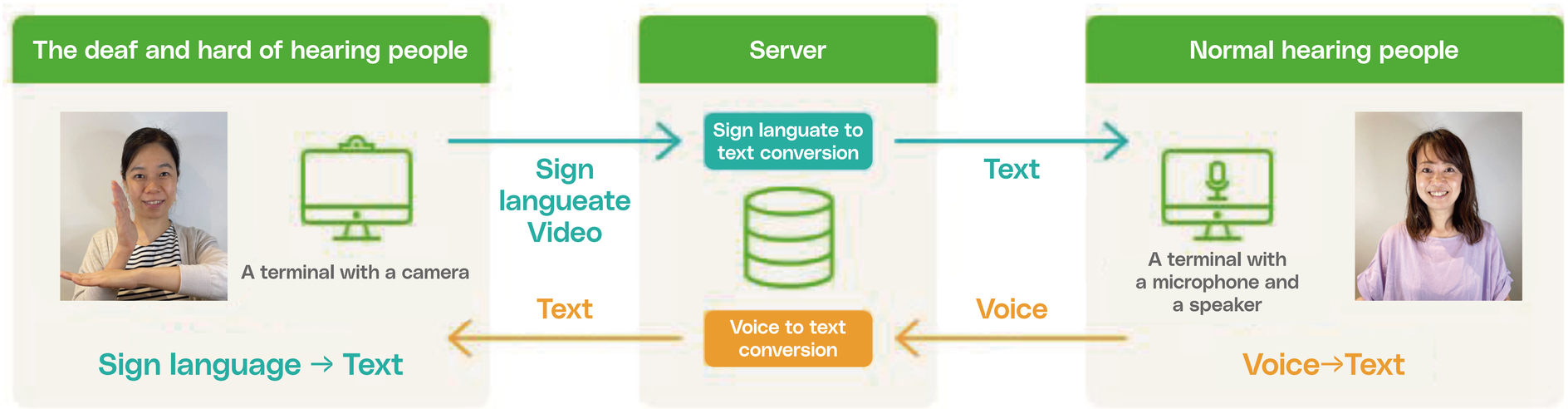
SoftBank

Interactive communication system between sign language and voice (SureTalk)

～System configuration and AI leveraged elemental technologies～

[Joint R&D]
The University of Electro-Communications | SoftBank Corp.

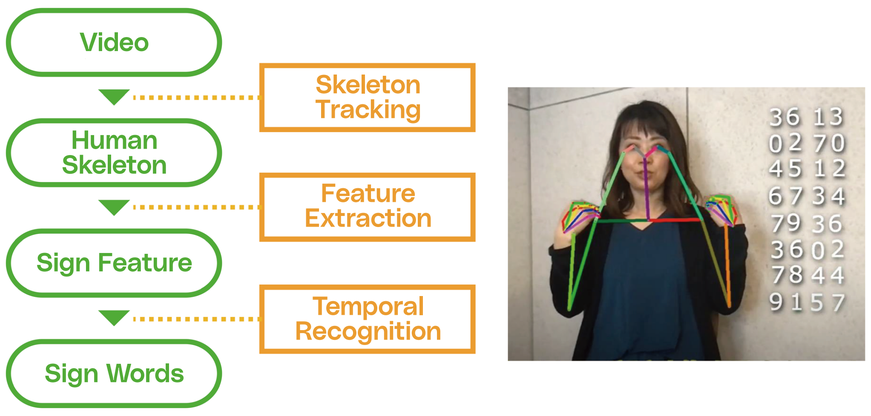
SureTalk System configuration



The most advanced AI technologies are utilized in order to create the structures needed in SureTalk.

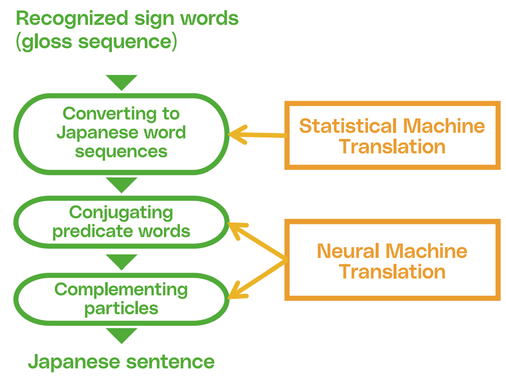
Continuous Sign Language Word Recognition

We have developed a video-based continuous sign language word recognition with skeleton tracking, sign feature extraction, and temporal recognition. First, the skeleton tracking extracts a human skeleton from a video. Next, the feature extraction extracts sign features which include human poses and motions. Finally, the temporal recognition outputs the continuous sign language words. It is available for in the wild without specific sensors and equipment because our approach is robust to the variation of human clothes and backgrounds. Moreover, we have successfully improved recognition performance using the latest deep learning techniques.
(Takahashi Hiroki Lab., The University of Electro-Communications)



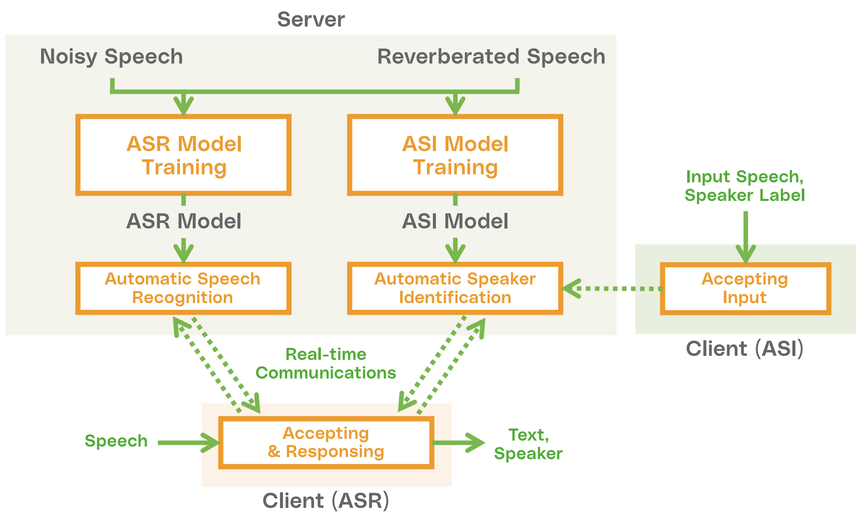
Translation from Signed Japanese glosses to Japanese Text

In the module of natural language processing, phrase-based statistical machine translation is used to map sign glosses to corresponding Japanese words or phrases, and then transformer-based neural machine translation is used to complement particles and conjugate predicates. This pipeline machine translation method successfully translates signed Japanese gloss sequences obtained by the video-based sign language word recognition into more natural Japanese sentences.
(Akira Utsumi Lab., The University of Electro-Communications)



Automatic speech recognition system for real environments

The speech processing module uses an open-source, deep learning-based speech recognition engine to perform real-time recognition of speech uttered by a person with normal hearing. We have independently collected speech data in a variety of environments, including noisy and reverberant environments, and used these data to train our acoustic model to improve recognition accuracy in real environments. We also aim to facilitate communication among several speakers by implementing a speaker identification system.
(Toru Nakashika Lab., The University of Electro-Communications)



国立大学法人 電気通信大学 The University of Electro-Communications

SoftBank