ALOFON - Metodyka i technologia polimodalnej alofonicznej transkrypcji mowy

Celem projektu było przeprowadzenie badań tworzących podstawy pod opracowywanie metod automatycznej transkrypcji fonetycznej mowy (w języku angielskim), opartej na wykorzystaniu połączenia informacji pochodzącej z analizy sygnałów fonicznych i wizyjnych. W szczególności, przeprowadzone zostały badania podstawowe nad związkiem pomiędzy zróżnicowaniem alofonicznym w mowie, tj. różnicami w charakterze tych samych głosek wynikających z różnego ułożenia artykulatorów w zależności od środowiska fonetycznego (tj. głosek sąsiadujących lub cech prozodycznych) a obiektywnymi parametrami sygnału. W ognisku uwagi badaczy znalazły się również parametry sygnału mowy (akustyczne i wizyjne) charakterystyczne dla Polaków uczących się języka angielskiego, w tym pozyskiwane przy wykorzystaniu systemu przechwytywania ruchów ust, pozwalającego uzyskać dodatkowe dane umożliwiające pogłębienie analiz odnoszących się do sposobu wymawiania głosek. Założeniem było opracowanie na tyle dokładnych metod analizy, żeby pozwalały one różnicować drobne zróżnicowania alofoniczne i akcentowe. W wyniku przeprowadzonych badań wykazano, że dzięki łącznej analizie sygnałów wizyjnych i fonicznych, transkrypcja fonetyczna mowy może zostać przeprowadzona z większą dokładnością, niż przy wykorzystaniu jedynie modalności akustycznej, tak jak opisano to we wcześniejszych pracach, dostępnych w literaturze. Pogłębione badania nad zróżnicowaniem głosek w kontekście parametrów sygnałów akustycznych i wizyjnych przyczyniły się do zaawansowania stanu wiedzy w dziedzinie audiowizualnego rozpoznawania mowy, a co za tym idzie w dziedzinie interakcji człowieka z komputerem. W toku badań zweryfikowano eksperymentalnie następujące hipotezy: 1. Łączna analiza danych fonicznych i wizyjnych poprawia skuteczność transkrypcji fonetycznej mowy na poziomie alofonicznym. 2. Przewidziana do opracowania metoda analizy sygnału mowy pozwala na pogłębioną w stosunku do wcześniejszego stanu wiedzy analizę zaawansowanych aspektów fonetycznych mowy. 3. Aspekty alofoniczne, takie jak m.in. nazalizacja, zaokrąglanie samogłosek, aspiracja, charakterystyczne cechy alofonów bocznych, mogą być skutecznie wykryte poprzez analizę sygnałów wizyjnych. 4. Różnice w sygnale mowy wynikające ze zróżnicowań alofonicznych i akcentowych mogą być zamodelowane z użyciem odpowiednich narzędzi matematycznych, a także rozpoznawane metodami uczenia maszynowego. Praktycznym wynikiem projektu jest ponadto opracowana i baza nagrań, dostępna pod adresem http://www.modality-corpus.org/. Całościowy korpus MODALITY składa się z ponad 30 godzin nagrań multimodalnych. Baza danych zawiera stereoskopowe strumienie wideo o wysokiej rozdzielczości, o wysokim klatkażu i sygnałów audio uzyskanych z macierzy mikrofonu i mikrofonu wewnętrznego, wbudowanego w komputer klasy notebook. Korpus można wykorzystać do opracowania systemu AVSR (audiowizualnego rozpoznawania mowy), ponieważ każda wypowiedź została ręcznie adnotowana. Opracowane rozszerzenie korpusu Modality, a mianowicie Alofon Corpus, zawiera dane audiowizualne i dane przechwytywania ruchu twarzy. Folder audio zawiera ścieżki dźwiękowe każdego nagrania zarówno dla mikrofonu kierunkowego, jak i mikrofonu krawatowego. Folder Vicon zawiera odpowiednie pliki dla każdego nagrania z kamer systemu przechwytywania ruchów mięśni twarzy.

Wyniki projektu upowszechniono w postaci 12 publikacji w czasopismach naukowych i 13 referatów konferencyjnych.

Projekt był finansowany przez narodowe Centrum Nauki, nr umowy 2015/17/B/ST6/01874.