وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

التعرف على لغة المتحدث في ملف صوتي باستخدام تقنيات ذكائية

رسالة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير علوم في
علوم الحاسوب

من قبل

محمد نايف عبد عبدالله

بإشراف

أ.د. فوزية محمود رمو

1443 هـ

2022 م

الملخص

تُعد اللغة أداة البشر لنقل الأفكار والاحاسيس وتمثل هوية الانسان حيث أصبحت اللغة اول هوية للجماعات في تاريخ البشرية وتمثل المخزون الثقافي وإرث الأجداد، يستخدم الافراد الصوت لبناء حلقة تواصل بين البشر ومن ثم التقارب بين الأمم والكلام هو لغة السمع.

تم بناء نظام حاسوبي ذكائي لكشف لغة المتحدث أطلق عليه اسم (Detection Speaker Language using Deep Learning (DSLDL) حيث يتكون هذا النظام من ثلاثة نماذج منفصلة، النموذج الأول للتعامل مع لغتين (العربية والإنكليزية) والنموذج الثاني للتعامل مع ثلاث لغات (العربية والإنكليزية والكردية) والنموذج الثالث للتعامل مع أربع لغات (العربية والإنكليزية والفرنسية والكردية) أطلق عليها 2L–DSLDL و 3L–DSLDL و 4L–DSLDL على التوالي.

يتكون نظام (DSLDL) المقترح من عدة مراحل، تم اولاً تهيئة قاعدة المعلومات لملفات الصوت حيث أُنشأت قاعدة معلومات لملفات صوتية سميت بـ(M2L–Dataset) وكانت محددة وغير محددة، فبالنسبة للغة العربية سُجلت عينات صوتية لمجموعة من الافراد البالغين (ذكور واناث)، ومن ثم تهيئة قاعدة المعلومات للغة الكردية حيث تم تجميع ملفات الصوت من الفيديوهات الخاصة بالدروس التعليمية، ومن ثم تجميع ملفات صوتية للغة الإنكليزية واللغة الفرنسية وبعد الانتهاء من هذه المرحلة تم اجراء المعالجة الأولية لجميع ملفات الصوت ثم استخلاص الصفات المهمة باستخدام خوارزمية معاملات درجة النغم (Mel Frequency Cepstral Coefficients) حيث تُعد هذه الخوارزمية من الخوارزميات الكفؤة عند التعامل مع ملفات صوتية وتم خزن المعاملات الناتجة بصيغتين اولاً ملف قاموس وثانياً صيغة صورة مخطط طيفي.

المرحلة الأخيرة هي مرحلة كشف لغة المتحدث باستخدام أكثر من طريقة ذكائية حيث تم استخدام الشبكة العصبية الالتفافية (Convolutional Neural Network) مع صور المخططات الطيفية وتم الحصول على أفضل نتيجة وكانت الدقة (98.40%)، واستخدام الشبكة العصبية الالتفافية (CNN) مع ملف القاموس وكانت الدقة (100%) وتم ايضاً استخدام شبكة الذاكرة طويلة

قصيرة المدى ثنائية الاتجاه (Bidirectional Long Short Term Memory) مع ملف القاموس وكانت الدقة (100%) بالنسبة للتنفيذ لنظام 2L-DSLDL للغتي (العربية والإنكليزية).

أما نظام 3L-DSLDL لثلاث لغات (العربية والإنكليزية والكردية) فقد تم الحصول على دقة بلغت (94.66%) لنموذج الشبكات العصبية الالتفافية (CNN) مع المخططات الطيفية، أما نموذج الشبكات العصبية الالتفافية (CNN) مع ملف القاموس فقد حصل على دقة بلغت (98%)، نموذج شبكة الذاكرة طويلة قصيرة المدى ثنائية الاتجاه (BiLSTM) حصل على دقة بلغت (99.19%).

بالنسبة لنظام 4L-DSLDL لأربع لغات (العربية والإنكليزية والفرنسية والكردية) فقد تم الحصول على دقة بلغت (92.79%) لنموذج الشبكات العصبية الالتفافية (CNN) مع المخططات الطيفية، أما نموذج الشبكات العصبية الالتفافية (CNN) مع ملف القاموس فقد حصل على دقة بلغت (97.29%)، نموذج شبكة الذاكرة طويلة قصيرة المدى ثنائية الاتجاه (BiLSTM) حصل على دقة بلغت (97.79%). كانت النتائج لنظام (DSLDL) المقترح جيدة جداً من حيث الدقة وسرعة التنفيذ.

Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Identification Speaker's Language in an Audio File Using Intelligent Techniques

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Computer Science**

By

Mohammed Naif Abd Abdullah

Supervised by

Prof. Dr. Fawziya Mahmood Ramo

2022 A.D. **1443 A.H.**

Abstract

Language is human's tool for transmitting the ideas and the feelings, and it represents the human identity, language formulated the first identity of groups in humanity history and represents the cultural potential and heritage of ancestors. People use the sound for building a communication link among humans and then rapprochement among nations and speech is the language of hearing.

An intelligent computer system was built to detect the speaker's language called (Detection Speaker Language using Deep Learning (DSLDL)), which consists of three separate models, first model for dealing with two languages (Arabic and English), second model for dealing with three languages (Arabic, English and Kurdish), and third model for dealing with four languages (Arabic, English, French and Kurdish) called 2L-DSLDL, 3L-DSLDL, and 4L-DSLDL respectively.

The proposed (DSLDL) system consists of several stages, first the database of audio files was adapted, and the researcher created a database of audio files called (M2L-Dataset) and it was customized and varied, as for the Arabic language, the researcher recorded audio samples for a group of mature individuals (male and female), and then adapted the database for the Kurdish language where the researcher has compiled the audio files from the videos of the educational lessons, and then he compiled audio files for the English and French languages, and after completing this stage, the preprocessing was carried out for all the audio files and then the important features were extracted using the Mel Frequency Cepstral Coefficients (MFCC). this algorithm is efficient when dealing with audio files, and the extracted features are stored in two forms, first: a dictionary file form, second: spectrogram form.

The last stage is detecting the speaker's language using more than one intelligent method, where the Convolutional Neural Network (CNN) was used with spectrograms, and the best result was obtained where the accuracy was (98.40%), and the Convolutional Neural Network (CNN) was used with the dictionary file and the accuracy was (100%), Bidirectional Long Short-Term Memory was also used with the dictionary file, and the accuracy was (100%) when dealing with both languages (Arabic and English).

As for the 3L-DSLDL system for three languages (Arabic, English and Kurdish), an accuracy of (94.66%) was obtained for the CNN model with spectral diagrams, while the CNN model with the dictionary file obtained an accuracy of (98%), the bi-directional long-term short-term memory (BiLSTM) network model obtained an accuracy of (99.19%).

For the 4L-DSLDL system for four languages (Arabic, English, French, and Kurdish), an accuracy of (92.79%) was obtained for the CNN model with spectral diagrams, while the CNN model with the dictionary file obtained an accuracy of up to (97.29%), the BiLSTM network model obtained an accuracy of (97.79%). The results of the proposed (DSLDL) system were very good in terms of accuracy and rapid implementation.