UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA

FACULTAD DE PRODUCCIÓN Y SERVICIOS ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

LA TESIS FORMATO ARTÍCULO TITULADO:

ISOLATED AUTOMATIC SPEECH RECOGNITION OF QUECHUA NUMBERS USING MFCC,

DTW AND KNN

RECONOCIMIENTO AUTOMÁTICO DE HABLA AISLADO DE NÚMEROS EN QUECHUA USANDO MFCC, DTW AND KNN

Presentada por el bachiller
HERNAN FAUSTINO CHACCA CHUCTAYA
Para optar el Título Profesional
de Ingeniero de Sistemas

Asesor:
Mgter. Eveling Gloria Castro Gutierrez

Arequipa - Perú

2019

Agradecimientos

- A mis padres, Octavio Chacca Zamata y Concepcion Chuctaya Huaracha, que siempre estuvieron ahí apoyándome para lograr mis objetivos, superar obstáculos y motivarme para llegar hasta este punto. Gracias porque cada una de sus enseñanzas, consejos y motivaciones hicieron que no me rindiera y siga adelante por más difícil sea el camino, que dedicara todo mi esfuerzo en lograr mis sueños y este trabajo es una paso más de muchos.

- A mi asesora Mgter. Eveling Gloria Castro Gutierrez, por permitirme contar con su conocimiento y capacidad de investigación que permitieron la realización de este trabajo haciendo fácil lo difícil. Gracias por toda la paciencia que tuvo para enseñar, corregir y guiar cada paso que dimos en el proceso para lograr este trabajo.

- Al Centro de Investigación CiTeSoft - UNSA (EC-0003-2017-UNSA), por proporcionarme materiales, libros y materiales para desarrollar mi investigación. Por darme un espacio para dedicar horas para estudiar e investigar de la cual el resultado es este trabajo.

- A toda mi familia que fueron un soporte moral y emocional a lo largo de mi vida. A mis amigos de la universidad por ser parte de estos cinco años en la cual compartimos experiencias y conocimiento. Por hacer que cada investigación, cada proyecto, cada tarea, clase sea amena, satisfactoria y divertida.

- A todas las personas que de una manera u otra, fueron claves en mi formación profesional. A todas las personas que ofrecieron su aporte para la realización de esta investigación. Y hacer una extensión especial a las personas que me prestaron su voz para grabar audios en quechua para la recolección de datos, mismo datos que fueron uno de los pilares para el desarrollo de esta investigación.

Resumen

El área de reconocimiento automática de voz(ASR) se define como la transformación de señales acústicas en palabras de cadena. Esta área se ha desarrollado durante muchos años para facilitar la vida de las personas, por lo que se implementó en varios idiomas. Sin embargo, el desarrollo de ASR en algunos idiomas con pocos recursos de base de datos pero con una gran población que habla estos idiomas es muy bajo. El desarrollo de ASR en el idioma quechua es casi nulo, lo que lleva a la cultura y la población a aislarse de la tecnología y la información. En este trabajo se desarrolla un sistema ASR de números Quechua aislados donde se implementan los métodos de Mel-Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW) y K-Nearest Neighbor (KNN) utilizando una base de datos compuesta por audios de números grabados en quechua desde uno hasta diez. Los audios grabados para alimentar la base de datos fueron grabados por hablantes nativos de quechua entre hombres y mujeres. La precisión de reconocimiento alcanzada en este trabajo de investigación fue de 91.1%.

Palabras Clave *— Reconocimiento automático de voz; MFCC; DTW; KNN..*

Abstract

The Automatic Speech (ASR) area is defined as the transformation of acoustic signals into string words. This area has been being developed for many year facilitating the lives of people so it was implemented in several languages. However, the development of ASR in some languages with few database resources but with a large population speaking these languages is very low. The development of ASR in Quechua language is almost null which leads culture and population isolation from technology and information. In this work an ASR system of isolated Quechua numbers is developed where Mel-Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW) and K-Nearest Neighbor (KNN) methods are implemented using a database composed by recorded audio numbers from one to ten in Quechua. The recorded audios to feed the data base were uttered by natives man and women speakers of Quechua. The recognition accuracy reached in this research work was 91.1%.

Keywords — Automatic Speech Recognition; MFCC; DTW; KNN.

Tabla de Contenidos

UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA FACULTAD DE INGENIERÍA DE PRODUCCIÓN Y SERVICIOS INGENIERÍA DE SISTEMAS

"Reconocimiento automático de Habla Quechua"

1. **Autores**
   - Autor
     Hernan Faustino Chacca Chuctaya
   - Coautor
     Rolfy Nixon Montufar Mercado
     Jeyson Jesus Gonzales Gaona

2. **Asesor**

Asesor N° 1	Eveling Gloria Castro Gutierrez
Grado académico:	Magister en Ingeniería de Software - UCSM
Institución de afiliación:	Universidad Nacional de San Agustín de Arequipa
Reseña del autor:	Investigador Calificado - Concytec Nº17735. Estudios de doctorado en Cs Computación en la Cátedra Concytec- Universidad Nacional de San Agustín de Arequipa (UNSA). Realizó una estancia tecnológica en la UChile, en el Departamento de Cs Computación (DCC) asesorada por la Dra. Nancy Hitschfeld, en el área de Visión Computacional con una beca otorgada por RDE No 037-2015-FONDECYT-DE (Abril- Julio 2015). Recibe el "RECONOCIMIENTO A LAS MUJERES CIENTIFICAS EN LA INVESTIGACION UNIVERSITARIA Y MUJERES DESTACADAS EN EL PERU", por la ANR (Octubre-2013). En agosto del 2013 recibe el grado de Magister en Ingeniería de Software de la Universidad Católica de Santa María(UCSM). Recibió el grado de Ingeniero de Sistemas de la UCSM (1995). Fue profesora visitante en Clayton State University (Atlanta - USA, 2007). Realizó una pasantía, como investigadora en Paralelismo, en el Laboratoire LE2I (UMR - CNRS, 2010), en la Universidad de Bourgogne, Dijon-Francia. Fue Directora de la Escuela Profesional de Ingeniería de Sistemas Período 2011-2013. Actualmente enseña en la Escuela Profesional de Ingeniería de Sistemas - UNSA y en la Escuela Profesional de Ingeniería de Sistemas - UCSM. En el año 2011-2012 participó como docente investigador en el proyecto MediMercado, No 068-FINCyT-PITEI, y participó en el proyecto No 128- FINCyT-FIDECOM-PIPEI-2012 denominado

	ASISTENCIA Y MONITOREO REMOTO DE PACIENTES POR MEDIO DE SENSORES, DISPOSITIVOS MÓVILES Y CENTRAL MEDICA VIRTUAL, en convenio con la empresa Microdata SRL. Se encuentra como Investigador Principal en Proyectos en el área de Visión computacional, Videojuegos, Lenguajes de Programación, Semántica del Lenguaje, Ha publicado artículos de Investigación en SCOPUS y Web of Science. Ha formulado proyectos de investigación desde el 2015. *h*-index: 2.

3. Planteamiento del problema

El reconocimiento automático de voz(ASR) es un técnica que permite la traducción del habla en un texto escrito. Entre las varias aplicaciones que se estudian está el desarrollo de sistemas de reconocimiento de voz de idiomas desconocidos o con idiomas con pocos recursos para su desarrollo [13][15][16]. El desarrollo de sistemas ASR se divide en general en dos fases: La fase de entrenamiento y la fase de reconocimiento. La fase de entrenamiento consta de tres etapas generales: a) La etapa de preprocesamiento en la cual se hace la segmentación de audios grabados. b) La etapa análisis en la cual se genera el modelo acústico y el modelo del lenguaje [12]. Se puede desarrollar un modelo Hidden Markov Model (HMM) para el modelo acústico y un modelo probabilístico del lenguaje para el modelo del lenguaje ya que el idioma quechua tiene un vocabulario amplio. c) La etapa de reconocimiento donde ya el sistema reconoce un discurso para transcribirlo a texto. Sin embargo, el principal reto para esta investigación es el dataset del idioma, que básicamente son el vocabulario completo del idioma y audios grabados para la fase de entrenamiento. A partir del análisis del idioma se definen las técnicas y modelos adecuados para lograr la mas alta precisión de reconocimiento [14].

4. Objetivo General

- Desarrollar un sistema de reconocimiento automático de voz del idioma quechua para su preservación a través de textos escritos..

5. Objetivos Específicos.

- Recopilar y analizar el estado del arte.
- Crear un dataset del idioma Quechua.
- Efectuar el reconocimiento automático de números.
- Efectuar el reconocimiento automático de de oraciones.
- Establecer textos escritos de quechua como repositorio de literatura y aprendizaje.

6. Descripción del proyecto

Según [1] entre 7 y 8 millones de personas hablan quechua actualmente en países de sudamericana. La gran mayoría de esta persona están concentradas en Ecuador, Perú y Bolivia. En Perú el 13.2% de la población, aproximadamente 3 millones, declararon saber al menos una variante del quechua.

Analizando el perfil de un quechua-hablante en [2], el 66% de la población quechua-hablante supera los 40 años de edad, un 23% tiene entre 25-39 años, y tan solo un 11 % tiene 18-24 años. Estos datos indican dos hechos muy críticos si los datos no cambian de manera positiva en los siguiente años; se perderá un 66 % de la población quechua hablante en el transcurso de pocos años, y que solo el 34 % de la población intentará mantener el quechua en nuestro país, considerando dentro de este ultimo grupo a la población quechua-hablante que se encuentra entre 18 y 39 años de edad. Para mantener este idioma se puede puede aprovechar que se cuenta con una gran cantidad de nativos tecnológicos en la población y utilizar nuevas tecnologías disponibles en la actualidad como el reconocimiento automático de voz y sus aplicaciones.

El reconocimiento de voz es cada vez mas utilizado con fines educativos, recuperación de lenguas desconocidas en proceso de desaparición y acceso a información por lo que es un medio con el cual se puede contribuir a la preservación del idioma quechua [3][4][5][6][7].

Una forma de preservar un idioma es promover su literatura manteniendo textos escritos que puedan actuar como repositorios [8]. Dado que mas del 20% de la población quechua-hablante no sabe escribir ni leer quechua pero si hablar [9], se puede emplear un sistema de reconocimiento automático de voz para plasmar textos escritos de la literatura quechua o algun material de aprendizaje. Esto puede darse con la contribución de personas que pertenecen a esta fracción de población que tiene la habilidad de hablar el quechua pero sin tener posibilidad que plasmarlo en un texto escrito.

7. Justificación

En Perú solo el 13.2% de la población, aproximadamente 3 millones, declararon saber al menos una variante del quechua [1].
Analizando el perfil de un quechua-hablante en [2], el 66% de la población quechua-hablante supera los 40 años de edad, un 23% tiene entre 25-39 años, y tan solo un 11 % tiene 18-24 años. Estos datos indican dos hechos muy críticos si los datos no cambian de manera positiva en los siguiente años; se perderá un 66 % de la población quechua hablante en el transcurso de pocos años, y que solo el 34 % de la población intentará mantener el quechua en nuestro país, considerando dentro de este ultimo

grupo a la población quechua-hablante que se encuentra entre 18 y 39 años de edad. Para mantener este idioma se puede puede aprovechar que se cuenta con una gran cantidad de nativos tecnológicos en la población y utilizar nuevas tecnologías disponibles en la actualidad como el reconocimiento automático de voz y sus aplicaciones.

Debido que a que más del 20% de los quechua-hablantes en Perú no escribe ni lee quechua y que la decreciente población quechua-hablante puede afectar a la preservación de la conocimiento y la herencia cultural e histórica que se esconde en el idioma quechua, la presente investigación pretende contribuir a la preservación del idioma quechua en nuestro país.

8. Delimitación

○ Delimitación Geográfica

El reconocimiento automático de voz se construye a partir de una estructura del idioma y un modelo acústico del mismo [10]. El quechua tiene una variedad de dialectos y se delimitara al grupo IIC de la clasificación de la familia de idioma quechua presentado por [11] que corresponden a los dialectos hablados en los de- partamentos Ayacucho, Apurimac, Cusco y Arequipa. Sin embargo, la variación dentro del grupo II-C es aun amplia por lo que el contexto que se desarrolla esta investigación se delimita al quechua de la región hablada en la region de Arequipa.

9. Referencias

[1] P. C. Rivera and L. G. Barron, “El quechua y sus hablantes: En la Pontificia Universidad Católica del Perú,”

[2] GFK, “quechua Statistic,” tech. rep., 2015.

[3] H. Prakoso, R. Ferdiana, and R. Hartanto, “Indonesian Automatic Speech Recognition System Using CMUSphinx Toolkit and Limited Dataset,” in International Symposium on Electronics and Smart Devices (ISESD) November 29-30, 2016, 2016.

[4] P. Wani, U. G. Patil, S. Bormane, S. D. Shirbahadurkar, and D. Y. Patil, “Automatic Speech Recognition of Isolated Words in Hindi Language,”

[5] M. Plauche, U. Nallasamy, J. Pal, C. Wooters, and D. Ramachandran, “Speech Recognition for Illiterate Access to Information and Technology,”

[6] M. Kumar, N. Rajput, and A. Verma, “A large-vocabulary continuous speech recognition system for Hindi,”

[7] M. Wald, “Using Automatic Speech Recognition to Enhance Education for All Students: Turning a Vision into Reality,” 2004.

[8] L. Page, “Nuevas formas de evitar la extinción de las lenguas,” 2016.

[9] J. Edgar Vargas Muñoz, J. Antonio Cruz Tello, and R. Alexander Castro Mamani, “Let’s Speak Quechua: The Implementation of a Text-to-Speech System for the Incas’ Language,” 2012.

[10] A. V. Anand, P. Devi, J. Stephen, and B. V. K, “Malayalam Speech Recognition System and Its Application for visually impaired people,”

[11] “La familia lingüística quechua,”

[12]M. A. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. M. d. Li- berto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E. C. Lalor, N. F. Chen, P. Hager, T. Kekona, R. Sloan, and A. K. C. Lee, "ASR for Under-Resourced Languages From Probabilistic Transcription," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, pp. 50–63, 1 2017.

[13] F. D. Rahman, N. Mohamed, M. B. Mustafa, and S. S. Salim, "Automatic speech recognition system for Malay speaking children: Automatic speech recognition system," in Proceedings of the 2014 3rd ICT International Senior Project Conference, ICT-ISPC 2014, 2014.

[14]M. Y. Tachbelie and S. T. Abate, "Effect of Language Resources on Automatic Speech Recognition for Amharic".

[15]J. Winebarger, S. S. Uker, and A. Waibel, "Adapting Automatic Speech Recognition for Foreign Language Learners in a Serious Game," 2014.

[16] M. Ali, M. Elshafei, M. Al-Ghamdi, H. Al-Muhtaseb, and A. Al-Najjar, "Generation of arabic phonetic dictionaries for speech recognition," in 2008 International Conference on Innovations in Information Technology, IIT 2008, 2008.

Isolated Automatic Speech Recognition of Quechua Numbers using MFCC, DTW and KNN

Hernan Faustino Chacca Chuctaya
National University of San Agustin
Arequipa, Peru

Rolfy Nixon Montufar Mercado
National University of San Agustin
Arequipa, Peru

Jeyson Jesus Gonzales Gaona
National University of San Agustin
Arequipa, Peru

Abstract—The Automatic Speech (ASR) area is defined as the transformation of acoustic signals into string words. This area has been being developed for many year facilitating the lives of people so it was implemented in several languages. However, the development of ASR in some languages with few database resources but with a large population speaking these languages is very low. The development of ASR in Quechua language is almost null which leads culture and population isolation from technology and information. In this work an ASR system of isolated Quechua numbers is developed where Mel-Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW) and K-Nearest Neighbor (KNN) methods are implemented using a database composed by recorded audio numbers from one to ten in Quechua. The recorded audios to feed the data base were uttered by natives man and women speakers of Quechua. The recognition accuracy reached in this research work was 91.1%.

Keywords—Automatic Speech Recognition; MFCC; DTW; KNN

I. Introduction

Technology has been facilitating people's lives since it became an integral part of their lives. It makes the communication with computers easy and one of the ways to do that is emulating human intelligence to understand what a person says aloud. [1]. The interaction between a person and a computer using the voice becomes simpler and more comfortable because it does not need special skills such as hand coordination and speed when typing with a keyboard
[2]. For this reason ASR systems were developed in many languages, including languages that have few resources in database, with the aim of making people's interaction with computers easy and thereby facilitating access to information. and technology [3] [4] [5].

ASR is the area of artificial intelligence that transform the audio signals spoken by a person into a sequence of words that can be understood for a computer [6]. It has been researched for years how a person can communicate with a computer in the same way a person communicates with another person [7]. The development of ASR covers issues from research on voice recognition to the implementation of dictionaries based on the speech spoken by a person and all of these issues divide the ASR into three types, ASR: from isolated words, continuous and connected words [8].

ASR systems of isolated words take as input individual words or a list of words with a well defined pause between them and each of the words is processed individually [1]. In this research work an ASR system is developed for isolated words, to be precise, for natural numbers from one to ten in Quechua, which is an official language in Peru.

Quechua is essentially an agglutinative language and this peculiarity makes Quechua different from the rest of the dominant languages in South America, thus this language is suffering a strong social pressure [9]. This goes hand in hand with the fact that the development of technology in these languages is very low which leads to the isolation of Quechua-speaking people from information and technology. The devel-opment of an ASR system in Quechua will enable people who speak only this language to use the technology to greater extent without the knowledge of operating with computer keyboard developed in foreign language and understanding information published also in foreign language.

This research paper presents the development of an ASR system of isolated words having a limited database. The rest of this research is organized as follows: Section II describes a review of the works related to this work. Section III provides the theoretical framework of ASR. Section IV develops the methodology used to implement the ASR system that this work proposes. Section V analyzes the results obtained from the ASR system and finally in section VI summarizes the conclusions reached through the development of this work.

II. Related Works

Atif in [10] developed a system for automatic recognition of isolated words with English language. In the phase of extraction of characteristics of an audio, MFCC was used and DTW and KNN were used in the recognition and classification block. DTW to make match the features of different audios and KNN to classify taking the characteristics that more resemble. The audios used were acoustically balanced and free of ambient noise. The recognition accuracy achieved in the work of these authors is 98.4%.

Wani in [2] developed an automatic recognition system for isolated words with the Hindi language. It is taken into account that many people who speak this language can not speak English, which is the language which the ASR systems were most developed with, and they can not access easily to this technology. For feature extraction, MFCC technique was used, and KNN and GMM (Gaussian Mixture Model) were implemented in the recognition phase. In order for the system to be independent of the speaker, the training audios of different speakers between men and women were obtained. Wani's work reaches a recognition accuracy of 94.31%.

In Indonesia, an ASR was developed using a tool based on HMM and with a limited database. [11] needed to build an acoustic model, a language model and a dictionary to develop the ASR for the Indonesian language. Own models of the numbers were developed which were used as input for CMUSphinx toolkit, which is the tool they used. The use of acoustic models already implemented to evaluate them under different SNR conditions was also investigated. The best recognition accuracy achieved is 86% and by experiment different noise level conditions the best accuracy is 80%.

Anand in [5] developed a modern ASR of wide vocabulary with an application in people with visual disabilities. In feature extraction phase, MFCC was used and in the classification and recognition phase an acoustic model was developed using thirty hours of HMM-based audio. To handle pronuncia-tion variation, a hybrid model was used between rule-based methods and statistical methods. The audio recordings were collected from 80 native speakers of the Malay language. The best recognition accuracy achieved is 80% and the developed system was integrated into OpenOffice Writter as a text entry interface through voice.

On the other hand, Ranjan develops an ASR system for isolated words from a language dialect called Maithili [12]. To obtain the necessary acoustic vectors for classification, the author implements MFCC. The system developed by Ranjan is an ASR system based on the HMM model. The acoustic model and the language model are developed with HMM. The recognition accuracy reached in the work described is 95%. However, future work is planned to improve accuracy in noise environments.

Speech recognition for people with amputated vocal cords differs to some degree from a common ASR. While it is true that the duration and intonation of words and vowels are practically the same, the pre-processing of the signals must be deeper. This problem is contemplated and developed by Malathi in [13] using MFCC for feature extraction of the audios and thus built the acoustic vectors. The classification or recognition was developed with GMM and Gradient Descent Radial Basis Function (RBF) Networks. The learning rate of the network are made proportional to probabilities density obtained from GMM. The result of the research was applied to patients who pronounced words only with the esophagus.

Bhardwaj [14] developed three schemes or types of ASR with the same methodology to evaluate the behavior of this methodology in different contexts. The types of ASR that are evaluated are: dependent on the speaker, multi speaker, and independent of the speaker. The methodology used starts by implementing MFCC for feature extraction of the audio. The acoustic model and the language model are based on HMM. To classify the words in Hindi, the language which they worked with, they used the K-Mean algorithm. The recognition rate for the independent speaker ASR was 99%, for the multi-speaker it was 98%, while for the independent speaker ASR it was 97.5%.

Ananthi developed an ASR for people with hearing prob-lems [15]. If the words of an announcer are interpreted by the computer and are simultaneously transcribed into text, a person with hearing impairment can easily understand any person. An ASR of isolated words based on HMM is developed in Ananthi's work. Because the focus of the work we are describing is aimed to the use of ASR in a fluent conversation, the implementation of DWT is discarded since it only works properly in isolated word ASR. The result of this work was successfully implanted in a population of people with hearing problems.

III. ASR

ASR systems are composed of two main blocks, a feature extraction block and a classification block [10]. The feature extraction block obtains values from an audio and these are passed to the classification block that is responsible for predicting the word or sequence of words corresponding to the input audio [16].

To express the audio signals in numeric values, there are a lot of algorithms and methods in feature extraction block. Some of these methods are: Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Cepstral Analysis, Mel-Frequency Scale Analysis, Filter-Bank Analysis, Mel-Frequency Cepstrum Co-efficients (MFCC), Kernal Based Feature Extraction, Dynamic Feature Extraction, Wavelet based features, Spectral Subtraction and Cepstral Mean Subtraction (CMS) [17]. According the review we made of related works, the most common and appropriate methods used in this type of ASR for feature extraction are MFCC and LPC and in this research work, the method we used is the first one, MFCC.

In the classification block, there are two main components, the acoustic model and the language model [18]. The acoustic model models how the pronunciation of a word is represented, and on the other hand, the language model models the prob-ability that a word fits a sequence of words. Hidden Markov model (HMM) and Neural Networks are the most common techniques for modeling an acoustic model and N-Gram model to model a language model. These techniques are common in continuous and wide vocabulary ASR [18] [19] [11] [20]. However, for ASR of isolated words and with a limited data set, there are techniques that behave better in these cases. In this type of ASR, only acoustic model is built to classify the words and the techniques like Dynamic Time Warping (DTW) and K-Nearest Neighbor (KNN) are the ones which reach better results to find similarity between the signals of two or more audios [10] [2].

After analyzing the architecture of a conventional ASR and ASR of isolated words ASR, the ASR that will be developed in the present work adapts the architecture of the ASR of isolated words that is constituted of two main blocks, which is the block of feature extraction and the block of classification, and in each block the algorithms that best adapt to our problem are implemented according to the state of the art review. The blocks of the architecture as well as the algorithms to be used are presented in Fig.1.

IV. METHODOLOGY

The ASR of isolated words that is developed in this work implements the MFCC technique for feature extraction. To classify the representation of the audio signals that MFCC provides, the DTW and KNN techniques are used. Before start

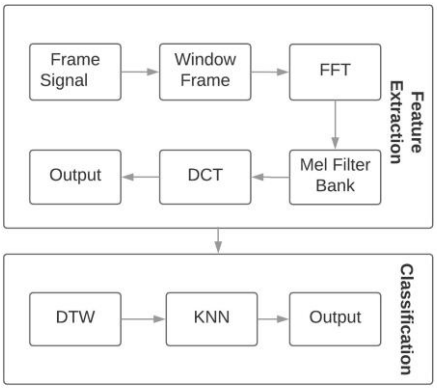

Fig. 1. Architecture of implementation of ASR.

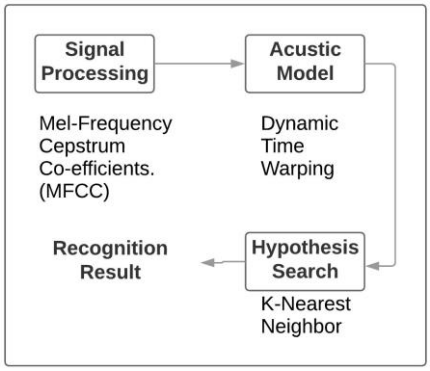

Fig. 2. Methodology of implementation of ASR Quechua.

the described process, the input audios go through a noise reduction and elimination filter. This methodology can be seen in Fig. 2.

A. Database

As a first step to implement this ASR, a Quechua database was developed. For them, isolated words from different native speakers were recorded. The numbers from 1 to 10 were obtained by recording thirty people between men and women, fifteen men and fifteen women. The numbers uttered and recorded can be seen in Table I, and each number was saved in an audio file with .wav. Each number was spoken by the thirty people, so in total we had three hundred audio to be processed and put them in the ASR system we implemented.

B. Feature Extraction.

In this stage, MFCC is implemented and it is considered the most important stage where parametric representation of the audio signals determine how effective is the performance of the next stage, which is classification. MFCC is based on human auditory perception that can not perceive frequencies above 1000 Hz [21][22], in other words it is based on known variation of the human ears critical bandwidth with frequency . The best representation of these audio signals is the Mel scale, which is approximately linearly below the 1000Hz frequency

TABLE I. NUMBERS IN QUECHUA

One	Uc
Two	Iskay
Three	Kimsa
Four	Tawa
Five	Pisqa
Six	Soqta
Seven	Qanchis
Eight	Pusaq
Nine	Isqon
Ten	Chunka

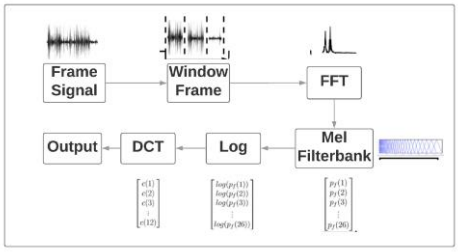

Fig. 3. Mel-Frequency Cepstrum Co-efficients.

and logarithmically above. The entire MFCC process can be seen in Fig. 3 and then each of the phases is developed.

1) Audio Pre-Processing.: Because we needed the acoustic vectors with the same longitude, we had to edit every audio's duration in order to have them with exactly one second of duration. These numbers were spoken in an acoustically balanced and noise free environment, thus it was not necessary to used any noise reduction technique. Every recording was saved in .wav format of 16-bit PMC and 8000Hz frequency. The signal obtained after pre-processing an audio can be seen in time series in Fig 4.

2) Pre-Emphasis.: We apply pre-emphasis to the original signal to amplify the high frequencies. According to [23] the pre-emphasis filter can be used in several ways: a) It balances the frequency spectrum since high frequencies usually have lower magnitudes than those of high frequencies. b) Avoid numerical problems during the operations of Fourier transformations. c) You can also improve the Signal-to-Noise Ratio (SNR). This filter is applied to a signal x using (1).

$$y(t) = x(t) - \text{↵}(t-1) \quad (1)$$

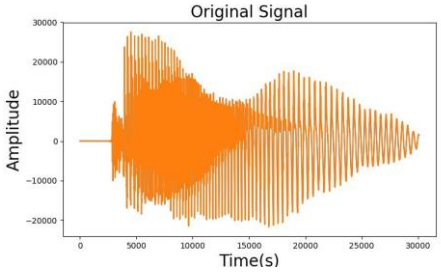

Fig. 4. Original Signal.

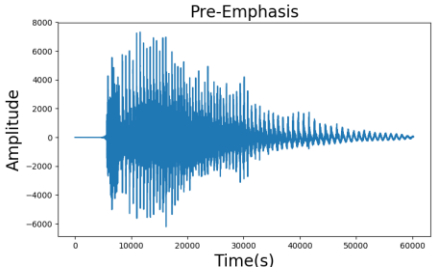

Fig. 5. Pre-Emphasis

After applying the pre-emphasis filter to the original signal, a new signal is shown, which can be seen in Fig 5. We can see that the amplitude of high frequency bands was increased and the amplitudes of lower bands was decreased so it will help to get slightly better results.

3) Framing and Windowing.: With the signal obtained from the pre-emphasis filter, a process is done in which the signal is divided into small frames, and this process is called framing. The reason for doing this process is that when doing the fourier transformation, which is the next step, you lose frequency contours if you work on the entire signal.

After dividing the signal into frames overlapped with each other, a Window function is applied to each frame to remove discontinuities and in this work the Hamming function is used. In this work, the Hamming function is used to counter the assumption made by Fast Fourier Transform (FFT) that the data is infinite and to reduce the spectral leak [23]. The equation of the Hamming function that is applied to each frame is described in (2) where "n" is the total number of samples in a single frame.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (2)$$

After applying the Hamming function, the output signal is plotted as shown in Fig. 6.

4) Fast Fourier Transform.: FFT is applied to the signal obtained in the previous section to transform each frame of N samples from a time based domain to a frequency based domain [21]. In other words, a frequency spectrum is calculated where N is generally 256 or 512, and (3) is used to calculate this result. The output of the FFT method is shown in Fig. 7 where the domain of the signal is the frequency.

$$P = \frac{|FFT(x_i)|^2}{N} \quad (3)$$

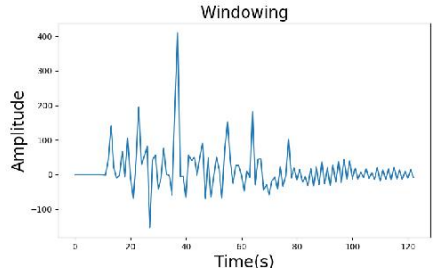

Fig. 6. Windowing

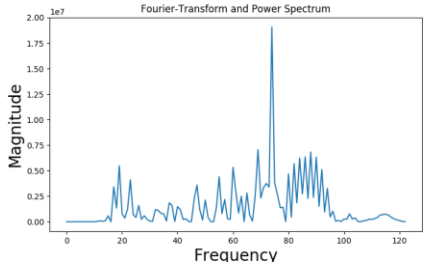

Fig. 7. Fast Fourier Transform

5) Mel Filter Bank.: The frequency range of the FFT spectrum is very wide and a voice signal does not follow a linear scale [21][24]. Filter Bank is then worked to transform the signal from Hertz to Mel scale as shown in Fig. 8 where the Mel filter bank comprises of triangular shaped overlapping filters. To calculate the filter banks, triangular filters are used, and the frequency in Hertz (f) can be converted to a Mel scale using (4).

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

6) Discrete Cousin Transform (DCT).: This is a process to convert the spectrum in Mel scale to a time-based domain. The result of this process is called MFCC. The set of coefficients obtained is called acoustic vectors [21]. In other words, until this phase, the inputs that were audios, are transformed into a sequence of acoustic vectors, which in turn, will form the set of inputs for the classification algorithms. The result is shown in Fig. 9.

C. Classification and Recognition

To evaluate the recognition accuracy, the development of the classification and recognition stage plays a very important role. In this work, DTW and KNN are used to find matches between different acoustic vectors obtained in the feature extraction phase. In DTW, the dynamic programming approach is used to find similarities between two time series, which basically have the same structure as the previously obtained acoustic vectors. For classification in continuous ASR is more accurate to use other techniques such as HMM or Neural

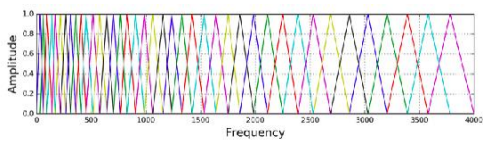

Fig. 8. Filter bank on a Mel-Scale

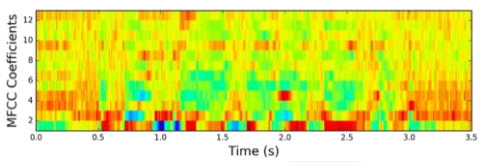

Fig. 9. MFCCs

Networks applied with different approaches such as Deep Learning [25]. These techniques are used because they try to imitate the human language learning taking into account variations of dialects or types of pronunciations. In this work an ASR of isolated words is worked so it is more appropriate to use DTW and KNN.

1) Dynamic Time Warping.: It is an algorithm to find the minimum distance between two sequences or time series dependent on certain values such as time scales, which was initially used only for ASR jobs but its application was extended to fields such as Data Mining [26][27]. Consider two time series P and Q with a length of n and m respectively.

P = p1, p2, p3, ... pn

Q = q1, q2, q3, ... qm

An mxn matrix is built and for each intersection the distance between the two points (pi, qj) is calculated using the Euclidean Distance formula described in (5).

$$d(p_i, q_j) = \sqrt{(p_i - q_j)^2} \quad (5)$$

Then the minimum accumulated distance is calculated using (6). DTW can have many variations with interesting improvements but each optimization is developed under a specific domain and it is difficult to use it in fields like ASR [28].

$$D(i, j) = \min[D(i-1, j-1), D(i, j-1), D(i-1, j)] + d(i, j) \quad (6)$$

2) K-Nearest Neighbor.: Given an ”n” point, K-Nearest Neighbor is an algorithm that finds all values closest to ”n” within a set of values that make up the training database [29]. In ASR, a feature vector takes the value of ”n”, and KNN finds the vectors closest to ”n” taking as reference a distance metric as the Euclidean distance that is calculated between all the vectors with the DTW algorithms.

V. Analysis of Results

The experiment was conducted on a database of three hundred natural number audios from one to ten in Quechua. Each audio in .wav format had exactly a duration of 60 seconds. Each number was pronounced by thirty different people, between men and women.

The database was divided into two sets, one for training that corresponds to 70% of the audios and another for the test that corresponds to 30% of the audios. Of the 90 numbers that passed the classification method, the number of correctly classified numbers was 82. Using (7) the accuracy of recogni-tion of the ASR developed in this research work is calculated, which at the end of the experiment reached a value of 91.1%.

$$\text{Accuracy} = \frac{\text{words detected correctly}}{\text{number of words in data set}} \quad (7)$$

The results were also analyzed in the form of a normalized confusion matrix where we can see more details of the

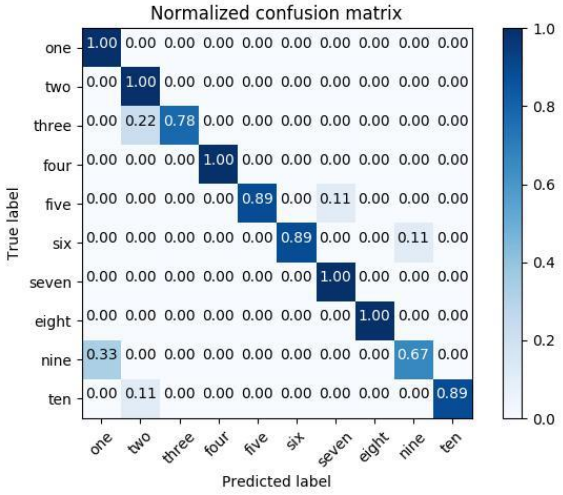

Fig. 10. Confusion Matrix

performance of the ASR system [30]. The confusion matrix for our system can be shown in Fig. 10 where the result of the classification is taken for each number. In the matrix, it is shown the accuracy of recognition for each label that is useful to analyze which numbers are correctly recognized or partially well recognized. Furthermore, we can identify the numbers with low recognition accuracy to analyze its features and improve the system to achieve a high recognition rate.

VI. Conclusion and Future Work

This research work develops an ASR system for isolated words using the MFCC, DTW and KNN techniques. The architecture in which we worked is divided into two blocks, feature extraction block and, the classification and recognition block. In each block we used algorithms that adapt better to our problem. In feature extraction block, it was developed implementing MFCC that consists of a series of algorithms that work sequentially. In the classification block, the acoustic vectors obtained in feature extraction block were classified using DTW and KNN. The results were evaluated using (7) which at the end of the work reached a value of 91.1%. The results were also analyzed in the form of a confusion matrix which shows us the recognition accuracy for every number to identify which numbers are the most recognized and which ones are partially recognized.

As future work it is proposed to improve the ASR system developed in this work including all the words that are spoken in Quechua. Next it is proposed to develop this system as a continuous speech recognition system capable of understand-ing and processing an speech spoken by any Quechua native people in a fluent way.

References

[1] S. Masood, M. Mehta, Namrata, and D. R. Rizvi, “Isolated word recognition using neural network,” 2015 Annual IEEE India Conference (INDICON), pp. 1–5, 2015.

[2] P. Wani, U. G. Patil, D. S. Bormane, and S. D. Shirbahadurkar, “Automatic speech recognition of isolated words in Hindi language,” in Proceedings - 2nd International Conference on Computing, Commu-nication, Control and Automation, ICCUBEA 2016, 2017.

[3] M. A. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. M. d. Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E. C. Lalor, N. F. Chen, P. Hager, T. Kekona, R. Sloan, and A. K. C. Lee, "ASR for Under-Resourced Languages From Probabilistic Tran-scription," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 50–63, 1 2017.

[4] C. Wei and Y. Yang, "Mandarin isolated words recognition method based on pitch contour," Proceedings - 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, IEEE CCIS 2012, vol. 1, pp. 143–147, 2013.

[5] A. V. Anand, P. S. Devi, J. Stephen, and V. K. Bhadran, "Malayalam Speech Recognition System and Its Application for visually impaired people," 2012 Annual IEEE India Conference (INDICON), pp. 619–624, 2012.

[6] D. OShaughness, "Automatic Speech Recognition," CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, 2015.

[7] O. L. A. M. Ali, H. A. Shedeed, M. F. Tolba, and M. Gadalla, "Morphame-Based Arabic Language Modeling for Automatic Speech Recognition," in Mathematical Applications in Science and Mechanics, 2013, pp. 152–157.

[8] S. Nur, A. Mohamad, A. A. Jamaludin, and K. Isa, "Speech Semantic Recognition System for an Assistive Robotic Application," IEEE International Conference on Automatic Control and Intelligent Systems, 2016.

[9] J. Edgar Vargas Munoz,˜ J. Antonio Cruz Tello, and R. Alexander Castro Mamani, "Let's Speak Quechua: The Implementation of a Text-to-Speech System for the Incas' Language," 2012. [Online]. Available: http://www.unsaac.edu.pe/

[10] "Isolated Word Automatic Speech Recognition (ASR) System using MFCC, DTW & KNN," The 2016 Asia Pacific Conference on Multimedia and Broadcasting, 2016.

[11] H. Prakoso, R. Ferdiana, and R. Hartanto, "Indonesian Automatic Speech Recognition System Using CMUSphinx Toolkit and Limited Dataset," in International Symposium on Electronics and Smart Devices (ISESD) November 29-30, 2016, 2016.

[12] R. Ranjan and R. Dubey, "Isolated Word Recognition using HMM for Maithili dialect," 2016 International Conference on Signal Processing and Communication, ICSC 2016, pp. 323–327, 2016.

[13] P. Malathi and G. R. Suresh, "Recognition of isolated words of esophageal speech using GMM and gradient descent RBF networks," 2014 International Conference on Communication and Network Tech-nologies, ICCNT 2014, vol. 2015-March, pp. 174–177, 2015.

[14] I. Bhardwaj and N. D. Londhe, "Hidden Markov Model Based Isolated Hindi Word Recognition," 2nd International Conference on Power, Control and Embedded Systems, 2012.

[15] S. Ananthi and P. Dhanalakshmi, "Speech Recognition System and Isolated Word Recognition based on Hidden Markov Model (HMM) for Hearing Impaired," International Journal of Computer Applications, vol. 73, no. 20, pp. 30–34, 2013.

[16] N. F. Chen and H. Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 12 2016, pp. 1–7.

[17] K. S and C. E, "A Review on Automatic Speech Recognition Architecture and Approaches," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 9, no. 4, pp. 393–404, 2016.

[18] M. Kumar, N. Rajput, and A. Verma, "A large-vocabulary continuous speech recognition system for Hindi,," IBM journal of research and development, vol. 48, no. 5.6, pp. 703–715, 2004.

[19] A. T. DeepaGupta, "Kannada Speech to Text Conversion Using CMU Sphinx," International Conference on Inventive Computation Technologies, 2016.

[20] M. Vikram, N. Sudhakar Reddy, and K. Madhavi, "Continuous Au-tomatic Speech Recognition System Using MapReduce Framework," IEEE 7th International Advance Computing Conference, 2017.

[21] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algo-rithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," Journal of Computing, vol. 2, no. 3, pp. 2151–9617, 2010.

[22] S. C. Sajjan and C. Vijaya, "Comparison of DTW and HMM for isolated word recognition," International Conference on Pattern Recognition, Informatics and Medical Engineering, PRIME 2012, no. 1, pp. 466– 470, 2012.

[23] H. Fayek, "Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between," 2016. [Online]. Available: http://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

[24] M. Najafian, W.-N. Hsu, A. Ali, and J. Glass, "Automatic Speech Recognition of Arabic Multi-Genre Broadcast Media," 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 353–359, 2017.

[25] F. D. Rahman, N. Mohamed, M. B. Mustafa, and S. S. Salim, "Automatic speech recognition system for Malay speaking children: Automatic speech recognition system," in Proceedings of the 2014 3rd ICT International Senior Project Conference, ICT-ISPC 2014, 2014.

[26] M. B. Lazreg, M. Goodwin, and O. C. Granmo, "Vector representation of non-standard spellings using dynamic time warping and a denoising autoencoder," 2017 IEEE Congress on Evolutionary Computation, CEC 2017 - Proceedings, pp. 1444–1450, 2017.

[27] Y. Lou, H. Ao, and Y. Dong, "Improvement of Dynamic Time Warping (DTW) algorithm," Proceedings - 14th International Symposium on Distributed Computing and Applications for Business, Engineering and Science, DCABES 2015, pp. 384–387, 2016.

[28] A. Sharabiani, H. Darabi, S. Member, A. Rezaei, S. Harford, H. John-son, and F. Karim, "Efficient Classification of Long Time Series by 3-D Dynamic Time Warping," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 47, pp. 1–16, 2017.

[29] I. Trabelsi, R. Amami, and N. Ellouze, "Automatic emotion recognition using generative and discriminative classifiers in the GMM mean space," 2nd International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2016, pp. 767–770, 2016.

[30] X. Kong, J.-Y. Choi, and S. Shattuck-Hufnagel, "Evaluating Automatic Speech Recognition Systems in Comparison With Human Perception Results Using Distinctive Feature Measures," pp. 5810–5814, 2017.