Voice Recognition System to Support Learning Platforms Oriented to People with Visual Disabilities

Ruben Gonzalez(✉), Johnnathan Muñoz, Julián Salazar, and Néstor Duque

Universidad Nacional de Colombia, Manizales, Colombia
{rdgonzalezb,johmunozmor,jnsalazarv,ndduqueme}@unal.edu.co

Abstract. The use of a speech recognition system allows access to an simple and efficient interaction, among others, for people with disabilities. In this article, an automatic speech recognition system is presented. It was developed as a system that allows an easy adaptation to different platforms. The model is described with clarity and detail looking for reproducibility by researchers who wish to resume and advance in this field. The model is divided into four stages: acquisition of the data, preprocessing, feature extraction and pattern recognition. Information concerning the functionality of the system is presented in the section named experiments and results. Finally, conclusions are set and a future work is proposed, in order to improve the efficiency and quality of the system.

Keywords: Voice command recognition · Mel Frequency Cepstral Coefficients · Assistive technology · Universal access to education

1 Introduction

As one of the most natural and comfortable modes of human interaction, speech could be considered an ideal medium for human computer interaction. A simple solution of this problem would be possible if machine could mimic the human production and understand of speech. One of the most promising applications of speech technology is spoken dialogue system, which offers the interface for simple, direct and hand free access to information [1]. This issue is particularly important if the user has a disability that prevents the standard interaction with the system.

However, Automatic Speech Recognition (ASR) is not a trivial problem [2]. The differences between speaker's speech, noise levels, variations in the speed of pronunciation and mood of the user, generate errors that affect the recognition task and produce very low success rates. In that sense, most of the current systems are restricted to controlled environments, uses limited to a small group of people or special requirements related to the microphone's positioning, resulting in unnatural interfaces [3].

Several methods for ASR have been proposed [3–5]. The most robust are based on Hidden Markov Models (HMM) [6]. Although currently these systems have achieved a high level of accuracy, many of them have still not succeeded in solving the problem of high computational cost. Moreover, most of them are presented as commercial software with high prices that prevent access to these solutions. Also, due to copyright, it

M. Antona and C. Stephanidis (Eds.): UAHCI 2016, Part III, LNCS 9739, pp. 65–72, 2016.
DOI: 10.1007/978-3-319-40238-3_7

is not possible to access to the source code, even if the objective is purely scientific, e.g., to implement improvements or couple such systems with new platforms.

There is thus a need to design an ASR system as own development, oriented to eliminate the limitations presented above, i.e., a simple system (low computational cost), accessible, reliable and suited to the requirements of any platform.

2 State of the Art

A systematic review of the state of the art for ASR systems was performed, using a tool named Tree of Science (TOS)[1], developed at the Universidad Nacional de Colombia. In this way, it was possible to take a look at the articles that have most influenced the progresses made by the scientific community in order to achieve improvements in terms of accessibility, accuracy and efficiency of ASR systems. In the first speech processing researches, it was used the Short Term Spectral Amplitude (STSA), using the minimum mean square error estimator (MMSE) [7]. By then, this algorithm was quite complex but offered greater accuracy compared to others. Later, the use of more robust methods based on Hidden Markov Models (HMM) began to grow. This technique uses the Mel-frequency Cepstral Coefficients (MFCC) during the stage of feature extraction [6]. Until today the HMM remain as an important approach for continuous speech recognition systems with large vocabulary, due to the good results that it produces. Moreover, other articles mentions a method known as PARADE (Periodic Component to Aperiodic Component Ratio-based Activity Detection), accompanied by a method of feature extraction called SPADE (Subband based Periodicity and aperiodicity Decomposition), which achieves significantly greater accuracy with regard to the final choice of the words [8]. In [9] it is used the Dynamic Time Warping algorithm (DTW), which is currently one of the most extensively used algorithms for not very complex systems because of the advantages that it offers in terms of the computational cost. However, the disadvantage of this method is that it is restricted to small vocabularies. Current research is focus on achieving the correct recognition of speech and also seeks to create and implement tools able to segment words, that is, identify the beginning and end of each word to reduce the complexity of processing required with continuous speech recognition [10].

3 Proposed Model

A system that allows identifying a given number of voice commands and which can be coupled, among many systems, with virtual learning tools is proposed. The possibility to interact with applications via voice can become a tool for inclusion, especially if that possibility has certain features such as ease of access and interaction by the end user and developers.

The proposed model can be extended for interaction with many applications, but it was conceived with the interest of providing the possibilities to interact with digital

[1] http://tos.manizales.unal.edu.co/.

learning resources to people with physical and sensory disabilities. Even though it already exist several tools like JAWS[2], the group decided to design an experimental tool that could be more easily coupled to other developments for diverse needs. In particular, it was incorporated to a framework named GAIATools, which is oriented to the construction of Accessible Learning Objects for inclusion of people with visual disabilities. The authoring tools that comprise GAIATools in its initial version are: dictionary, text editor and reader, game for learning and assessment through questionnaires. In addition, it guides the designer to develop learning objects and the visually impaired users to interact with them effectively in the context of educational activities [11].

Considering the context of developing countries, where socio-economic limitations are more evident, tools such as the proposed here are of particular importance. The motivation to target the audio recognition system primarily as a complement to educational tools born of the certainty that education is a fundamental way to overcome the above mentioned socio-economic problems.

Currently, the system is able to recognize ten isolated words in low noise environments. However, it is possible to increase the database by a simple process. The intention with this is that the system becomes adjustable to the particular needs of each user, who previously must record every new word. It is important to note that in its original design, the system recognizes words in Spanish, but there is no impediment if someone requires to enter a new command in another language to the database.

The proposed system is divided into four stages, as shown in Fig. 1. The first stage consists of the acquisition of the audio signal that contains the information that should be recognized. Subsequently, a preprocessing stage is executed, where the signal is filtered to eliminate noise and its unwanted segments, such as the silence at the beginning and end of the recording. Follows the stage of features extraction, where a matrix that contains the Mel Frequency Cepstral Coefficients (MFCC) is calculated. Finally, the system executes an algorithm to calculate a Euclidean distance that compares the features matrix entered related to the corresponding patterns stored in the database, this way making a decision that completes the process of recognition.

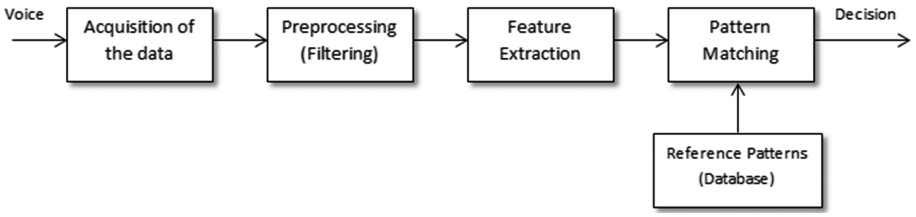

Fig. 1. Architecture of the proposed model

[2] http://www.freedomscientific.com/Products/Blindness/JAWS.

3.1 Audio Acquisition

During this stage, the user pronounces the word that is expected to be recognized. The system records a vector of t *Fs audio samples, where t is the time of recording and Fs is the sampling **frequency**; for this case 2 s and 44100 Hz, respectively. The sound is mono, i.e., only one audio channel is recorded.

3.2 Preprocessing

The preprocessing stage is divided, in turn, in three steps: filtering, normalization and silence suppression.

The implemented filter is the result of performing a process of spectral analysis on several audio recordings, where common noise frequencies in various environments (office, outdoor, home, etc.) were identified. Subsequently, a general frequency range where is expected to find useful information was established. Thus, the filter that provided the best results was a Hamming windowing FIR filter, band pass type, with sampling frequency of 44100 Hz and cutoff frequencies of 200 Hz and 8000 Hz.

Once the signal was filtered, it was normalized in order to limit it to standard values (between 0 and one) and thereby facilitate the development of the following steps. Normalization is especially important considering that the intention is to compare several audio samples that could possibly have different amplitude ranges, due to variables which control is difficult, such as the intensity of the speaker's voice or noise levels.

Finally, for silence detection and suppression, the signal is segmented and the energy of each segment is calculated. The value obtained must exceed a threshold to be considered as useful information, otherwise it will be deleted. The energy threshold was defined by comparing the noise energy values with the values obtained from randomly pronounced words. This is done in order to reduce the computational cost of the algorithm, which otherwise would have to consider signal segments that do not provide relevant information and instead may hinder the process of recognition.

3.3 Feature Extraction

The features used to represent each of the commands are the MFCC. According to the review of the state of the art, using these coefficients in an ASR system, it is possible to achieve an appropriate level of accuracy with low computational cost. The MFCC are a set of decorrelated parameters of the discrete cosine transform which are computed through a transformation of the logarithmically compressed filter-output energies, derived through a perceptually spaced triangular filter bank that processes the Discrete Fourier Transformed of the audio signal [12].

The MFCC were calculated as follows: the audio signal was segmented into intervals of 1024 samples length. The increase for segmentation is of 410 samples, which means that there is an overlap between the segments in order to avoid information lost when performing a windowing (through window Hamming in the time domain), considering that with this process the beginning and end of each segment will be attenuated. For each one of the segments 14 MFCC coefficients were calculated (excluding the 0th

coefficient). To perform this procedure 30 filters were used in the filter bank, with 0 as the low end of the lowest filter and 0.1815 as the high end of the highest filter. Once the MFCC are calculated, matrix of size NXC is stored, where N is the number of segments in which the audio signal was divided and C is the number of MFCC calculated, which are 14 for this case, as already mentioned. The number of rows of the matrices of features is not a constant because after performing the silence suppression in the recorded signals, the length of these may be different from the others. The parameters used for calculating the MFCC coefficients were selected according to the results obtained through several test, as the one described in the next section.

For each command that the system is able to recognize, it is necessary to calculate a matrix of features that will be stored as a "pattern matrix", and that represent the class of the command. Additionally, each time that a new recording is entered in order to execute the recognition; the system calculates a new matrix of features that will be known as "new matrix".

3.4 Decision Stage

The aim at this stage is to assign the entered matrix within a specific, in order to identify to which command corresponds the recording. This is accomplished by comparing the new matrix with the pattern matrices, looking for the best match among them.

To make the comparison between matrices, we must remember that each row represents a segment of the audio signal in terms of the MFCC coefficients. Therefore, to compare each row of the new matrix with the rows of the pattern matrix will be the same as comparing each segment of the new recording with the segments of the pattern recordings. However, there is no guarantee that the two recordings will match in terms of time of pronunciation, leading to compare different audible segments of the same word. In order to solve this problem, an individual error is calculated, represented by a Euclidean distance, between one row of the new matrix and each one of the rows of the pattern matrix. Later the least of these individual errors is taken and this value is assigned as the contribution of the row being analyzed to the total error. This process is repeated for each of the rows of the new matrix. At the end, there will be as many total errors as classes are taken and by identifying the least of them, it will be possible to know which command was pronounced in the entered recording. Although this process may seem complex, the operations are performed through algebraic arrangements rather than iterations, which reduce greatly the computational cost.

Finally, it has been established that the selected value as the minimum of the total error must be under a certain threshold; otherwise the user is requested to repeat the command more clearly in order to have an adequate level of reliability.

4 Experiments and Results

In order to test the performance of the implemented model, a test, which counted with the participation of 30 users, was designed. Since it was desired that the voices considered during the test were as dissimilar as possible, the participants were men and women

of various ages. The experiment, which was developed within moderate noise environments (study rooms, bedrooms, etc.), was divided into two parts: the first was the creation of a database with the voice of the participants and the second was the validation of the system.

The database was created trying that the recorded word was spoken out clearly and naturally. After performing this process, the participants were asked to repeat four times each of the following words in Spanish: open, back, enter, right, left, save, home, help, view, and internet; that is, for this stage 40 recordings were obtained for each user. In each attempt it was checked if the system was able to recognize the word or not. The detailed results of these tests can be seen in Table 1.

Table 1. Results of the test

User	Number of hits	Percentage
1	36	90 %
2	39	97.5 %
3	34	85 %
4	36	90 %
5	34	85 %
6	36	90 %
7	36	90 %
8	40	100 %
9	34	85 %
10	32	80 %
11	32	80 %
12	34	85 %
13	39	97.5 %
14	39	97.5 %
15	34	85 %
16	39	97.5 %
17	38	95 %
18	32	80 %
19	32	80 %
20	32	80 %
21	27	67.5 %
22	26	65 %
23	36	90 %
24	28	70 %
25	36	90 %
26	34	85 %
27	39	97.5 %
28	34	85 %
29	33	82.5 %
30	38	95 %

5 Conclusions

An ASR system has been proposed as a tool to complement the development of educational platforms, in order to make these affordable, among others, for people with visual disabilities. Although levels of recognition accuracy achieved are promising, the system does not reach levels as high as other commercial systems. However, the proposed model has a low computational cost, can be easily coupled with other platforms, allows the developer to make changes through a simple process and can be adapted to the particular needs of any user, allows for including new words in its database with a single recording.

As future work, the authors plan to consider other characterization methods such as autoregressive coefficients or strategies of classification as the Hidden Markov Models (HMM), in order to achieve continuous speech recognition, making even easier the interaction with applications. In addition, it is desirable to improve the preprocessing stage through the implementation of more efficient filters, this way achieving greater robustness in the system, eliminating the problems caused by differences in tone, pronunciation or noise. Finally, it is important to achieve the generalization of the system, in a way that it can be able to recognize any user without the need for prior registration of his voice.

Acknowledgments. The research presented in this paper was partially funded by the COLCIENCIAS project entitled: "RAIM: Implementación de un framework apoyado en tecnologías móviles y de realidad aumentada para entornos educativos ubicuos, adaptativos, accesibles e interactivos para todos (Implementation of a framework supported by mobile technologies and augmented reality for ubiquitous, adaptive, accessible and interactive learning environments for all)" of the Universidad Nacional de Colombia, with code 1119-569-34172.

References

1. Aggarwal, R.K., Dave, M.: Recent trends in speech recognition systems. In: Tiwary, U., Siddiqui, T. (eds.) Speech, Image, and Language Processing for Human Computer Interaction: Multi-modal Advancements, pp. 101–127 (2012)
2. Rosdi, F., Ainon, R.N.: Isolated Malay speech recognition using Hidden Markov Models. In: International Conference on Computer and Communication Engineering, 2008. ICCCE 2008, pp.721–725, 13–15 May 2008
3. Bedoya, W.A., Munoz, L.D.: Methodology for voice commands recognition using stochastic classifiers. In: 2012 XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), pp. 66–71, 12–14 Sept. 2012
4. Ibarra, J.P., Guerrero, H.B.: Identificación de comandos de voz utilizando LPC y algoritmos genéticos en Matlab. Rev. CINTEX **15**, 36–48 (2014)
5. Abushariah, A.A.M.; Gunawan, T.S.; Khalifa, O.O., Abushariah, M.A.M.: English digits speech recognition system based on Hidden Markov Models. In: 2010 International Conference on Computer and Communication Engineering (ICCCE), pp. 1–5, 11–12 May 2010
6. Gales, M., Young, S.: The application of hidden Markov models in speech recognition. Found. Trends Signal Process. **1**(3), 195–304 (2008)

7. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
8. Ishizuka, K., Nakatani, T., Fujimoto, M., Miyazaki, N.: Noise robust voice activity detection based on periodic to aperiodic component ratio. Speech Commun. **52**(1), 41–60 (2010)
9. Zhang, X., Sun, J., Luo, Z.: One-against-all weighted dynamic time warping for language-independent and speaker-dependent speech recognition in adverse conditions. PLoS ONE **9**(2), e85458 (2014). doi:10.1371/journal.pone.0085458
10. Komatani, K., Hotta, N., SATO, S., Nakano, M.: Posteriori restoration of turn-taking and ASR results for incorrectly segmented utterances. IEICE Trans. Inf. Syst. **E98.D**(11), 1923–1931 (2015)
11. Duque, N., Giraldo, M., Jaramillo, I.D., Salazar A.F.: GAIATools: Framework para la creación de objetos de aprendizaje accesibles. In: CAVA - VII Congreso Internacional de Ambientes Virtuales de Aprendizaje Adaptativos y Accesibles. Brasil (2015)
12. Hossan, M.A.; Memon, S.; Gregory, M.A.: A novel approach for MFCC feature extraction. In: 2010 4th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1–5, 13–15 Dec. 2010