The 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
21-23 September, 2017, Bucharest, Romania

Implementation of accent recognition methods subsystem for eLearning systems

Eugen Tverdokhleb[1], Hennadii Dobrovolskyi[2], Nataliya Keberle[2], Natalia Myronova[1]
[1] Software Engineering Department, Zaporizhzhya National Technical University, 64, Zhukovskogo str., Zaporizhzhya, Ukraine,
[2] Chair of Computer Science, Zaporizhzhya National University, 66, Zhukovskogo str., Zaporizhzhya, Ukraine
e-mail: {junta.kristobal, gen.dobr, nkeberle, natali.myronova}@gmail.com

Abstract — the results of the implementation of an external accent recognition system and its integration into massive open online courses platform Moodle are reported. Accent recognition becomes important in foreign languages learning to provide a feedback to a student on a presence of a certain unwanted accent in a foreign language pronunciation. Implementation of several accent recognition methods and their comparison is provided. It is shown that neural networks provide the most reliable recognition given the accented utterances from Wildcat Corpus of Native- and Foreign-Accented English.

Keywords — accent recognition; neural networks; machine learning; eLearning systems.

I. INTRODUCTION

The task of automatic foreign accent recognition is formulated as identification of the native language (L1) of a person speaking a second language (L2) [1].

The problem attracts great attention because common acoustic language models adjusted to match the "standard" native language corpus fail when applied to accented speech [2-6].

The accent detection problem is similar to spoken language identification, so the approaches to tackle the both tasks are tightly related.

The phonotactic-based methods utilize the accent-dependent sequences of the phonemes, and the acoustic methods estimate the differences in the phoneme statistics [7].

For instance, in the acoustic-based approach a spoken utterance is mapped to a sequence of short-time feature vectors which are assumed to have different distributions in different languages [8].

The simple way of taking into account the short phoneme sequence is the so-called shifted-delta-cepstral (SDC) features [9].

The long-term distinctive features of accents are modeled via a language- and speaker-independent universal background model (UBM) [10]. In [11] UBMs are used to extract low-dimensional i-vectors expressing voice recording of varying length as a single vector of fixed size keeping the language-specific information.

After the short-term or long-term features are extracted the machine learning methods [12] can be used to identify accents.

For instance the text-independent automatic accent classification system proposed by Angkititrakul and Hansen [13] is based on stochastic and parametric trajectory models.

Another popular approaches are support vector machines [8], hidden Markov models [14], long-short term memory neural networks [15], deep neural networks [16]. Accent recognition systems can also have application in eLearning to assist students learning foreign languages in improvement and check spelling.

The objective of this work is implementing accent recognition subsystem for one of the eLearning systems.

To reach this objective such sub-tasks must be solved:

- study the methods of speech sample acoustic features extraction;
- study the methods of speech sample classification based on extracted features;
- develop a method of the speaker accent recognition which will increase the reliability of recognition of certain accent;
- implement accent recognition methods and compare results;
- integrate the accent recognition system as a subsystem into one of the eLearning systems;
- test the implemented system using one of known corpora.

II. GENERAL CHARACTERISTICS OF SPEECH AND ACCENT RECOGNITION PROBLEMS

A. Speech recognition

Speech recognition addresses the problem of transforming a human speech into data. The first speech recognition device was created in 1952 [17]. Among the important applications of speech recognition are search engines development, voice controlled devices, digital personal assistants, teaching foreign languages.

Solving the problem of speech recognition can make possible creation of digital systems that receive input as spoken natural speech instead of typing commands or using graphical interface. This in many cases may simplify interaction with such digital systems, especially for the users who face difficulties using a keyboard or a touchscreen.

General procedure of speech recognition is as follows:

Step 1. Decide which parameters of the speech signal are needed, detect silence and voice.

Step 2. Parts of an audio signal containing speech are analyzed, audio features are extracted. Also syntactic and lexical analysis of speech sample is performed.

Step 3. Received parameters are compared to speech and acoustic models of known language to get the result of recognition process.

B. *Accent recognition*

Problem of accent recognition consists of a speech sample analysis and classification of the sample to attribute it to one of the known accents.

General procedure of accent recognition is as follows:

Step 1. Acquire a speech sample to classify.

Step 2. Analyze the sample, extract its acoustic features.

Step 3. Classify the sample: compare its features with the model of known accents for the given language.

III. METHODS OF SPEECH SAMPLE ACOUSTIC FEATURES EXTRACTION

The step of signal processing in the problem of recognition aims to extract information from the speech signal that has unique characteristic features of a particular accent. Extracted features are used for training the classifier or classification of the speech sample.

Generally accepted approach to speech signal processing is to use short-term analysis. That is, the signal is divided into fixed size short time periods called frames with the assumption that for a sufficiently short interval of a signal the characteristics remain the same. To analyze the speech signal the length of one frame is usually 10-30 ms, with overlaps between the frames equal to half of the length of the frame.

Then each frame can be processed with one of the following feature extraction methods: linear prediction cepstral coefficients [18] and Mel-frequency cepstral coefficients [19].

A. *Linear prediction cepstral coefficients (LPCC)*

The idea of linear prediction is that future values of a discrete signal can be approximated by the linear combination of previous samples from a certain number of observations [19].

The weight coefficients of such a linear combination are called linear prediction coefficients. Finding the coefficients of linear prediction is performed using the Durbin recursive algorithm [20].

Using the linear prediction coefficients it is possible to calculate cepstral coefficients that characterize the speech signal.

B. *Mel-frequency cepstral coefficients (MFCC)*

This method is the most common in speech recognition and speaker recognition systems and applies well to the accent recognition problem [19].

The input of the algorithm is the sequence of frames of the signal. It is subjected to the application of some weight function and Fourier transform. Weight function is used to reduce errors in Fourier analysis that are caused by finiteness of the sample.

The frequency range of the sample is converted to Mel scale. Mel scale is a result of research of human ear's ability to perceive sounds at different frequencies.

Filter bank is applied to the resulting signal after Fourier transforms, and then it is divided into ranges. Obtained values are transformed with a logarithm function. The final stage is a discrete cosine transformation.

The output of the algorithm is a vector of Mel-frequency cepstral coefficients for the given speech sample.

In practice, the MFCC method is the one of the most used because of its ease of use and better suitability for feature extraction of the speech signal to solve various problems related to speech recognition, including accent recognition.

IV. METHODS OF SPEECH SAMPLE CLASSIFICATION

The problem of speech signal recognition differs from the static pattern recognition, because in case of speech recognition the object of analyzing is a process and not a static image or pattern. So a speech sample to be recognized is represented by a series of feature vectors rather than a single one.

Several methods of classification exist to solve the recognition problem [21].

A. *Gaussian mixed models (GMM) method*

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs.

Modeling with mixtures uses a mix of several statistical distributions to model the distribution of a given population.

Gaussian distributions are used most often for mixture modeling [22].

In the area of speech recognition GMM is used for speech sample classification using its extracted acoustic features, many implementations and optimizations of the method exist.

B. *Hidden Markov models*

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. The problem in this case is to find the hidden parameters based on the ones, that are available for observation and influence the hidden ones [23].

In the area of speech recognition HMM can be considered as an improved Bayesian classifier for finding the most probable utterance from the raw speech sample.

Hidden Markov models better adapt to the dynamics of speech in terms of that the same phone can be pronounced differently depending on the sounds surrounding it in an utterance.

HMM acquired wide usage to solve the problem of speech recognition and transforming a raw speech signal into text using its acoustic features [24].

C. *Neural networks*

Recognition using neural networks is one of the perspective methods nowadays. Neural networks model recognition process similarly to one of biological systems using learning mechanisms.

For speech recognition fragments of a speech signal are passed as an input to a neural network classifier. The input layer of neurons has the same number of neurons as the number of features in each vector for the speech sample. The output is the resulting recognized text [25].

Neural networks are the most perspective method because the precision of systems, based on neural networks is very high due by the possibility to learn; also the speed of classification can be greatly improved by parallelization of computational processes, because neural networks are highly suitable for parallelization.

V. THE PROPOSED SPEAKER ACCENT RECOGNITION METHOD

In this work a combination of Mel-frequency cepstral coefficients method for feature extraction and neural network for classification was used to solve the problem of accent recognition.

MFCC was used because it allows to achieve good quality of feature coefficients vectors for an audio sample. Neural network classifier was used because of high precision of recognition that can be achieved on a given subset of accents.

At the first phase of recognition speech samples are processed by MFCC, which outputs a set of feature vectors. The size of the set is defined by the length of audio file, and the number of coefficients in each of the vectors is the same. So, the number of coefficients defines the number of input neurons in the classifier.

The number of output neurons equals the number of accents, represented in the dataset the model is created from.

VI. IMPLEMENTATION OF ACCENT RECOGNITION METHODS

Python programming language was chosen for implementing the system due to its popularity in data science and machine learning applications.

Feature extraction and classification methods described in previous sections are available in open-source libraries. For example, for feature detection scikits.talkbox library was chosen [26]. Scikits.talkbox – is one of the "toolkits" for a SciKits library. Scikits.talkbox is designed for audio files processing, mainly speech processing and feature extraction.

Scikits.talkbox library provides the following features:

- spectrum estimation related functions: both parametric, and non-parametric;
- Fourier-like transforms;
- basic signal processing tasks such as resampling
- speech related functionalities.

The library is under development, so more new features can be added.

For the purposes of accent recognition MFCC was used from the scikits.talkbox library for feature extraction and obtaining vectors of feature coefficient for speech samples.

A correspondent UML class diagram is presented in the Fig. 1. Classes that were developed in this work are inside the inner box on the diagram.

QMainWindow is a standard class for graphical interface creation from Qt library.

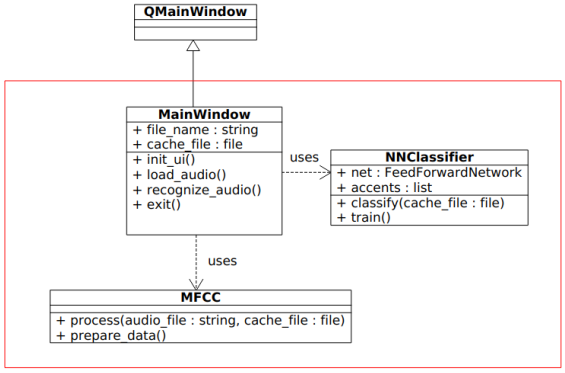

Figure 1. Class diagram of the developed program

Three classes were developed for the program: MainWindow, NNClassifier and MFCC.

MainWindow class has four methods:

- init_ui – this method is responsible for creation and initialization of the graphical interface of developed program;
- load_audio – this method is responsible for opening the dialogue window to choose a speech sample audio file and checks the file before processing;

- recognize_audio – this method is using corresponding methods of other developed classes to recognize the speech sample;
- exit – this method is called when the application is exited and is responsible for correctly closing of all used files and finishing the program.

MainWindow class represents the main application window and organizes interaction with the user.

MFCC class is a wrapper upon the tools of scikits.talkbox for feature extraction and saving features data of a speech sample to a temporary file for using it in the classification process.

The MFCC class has two methods:

- prepare_data – this method is responsible for the preparation of data samples; obtained feature vectors are saved to a file to use them for training of the classifier model;
- process – this method is responsible for processing and feature extraction of a given audio file of a speech sample using scikits.talkbox methods.

NNClassifier class implements a neural network classifier to recognize accent by feature vectors of a given speech sample.

This class has two methods:

- train – this method is responsible for building of the model based on the feature vectors of data samples; generated model is saved to and XML file to use it for classification;
- classify – this method gets speech sample feature vectors from a temporary file created my MFCC classes "process" method and uses them to classify speech sample.

VII. The Experimental Study of the Proposed Method

For the experimental study of the proposed method Wildcat Corpus of Native- and Foreign-Accented English was used [27]. The corpus contains 1342 audio recordings, representing these accents of English: proper English pronunciation; Korean; Chinese; Japanese; Italian; Indian; Irish.

The whole audio sample set was divided into training and testing subsets at a ratio of 4 to 1. The model of artificial neural network using backpropagation was built and then tested using the dataset.

The results of testing of three different speaker accent recognition methods are given in the Table 1. The following abbreviations are used in the table:

MFCC-NN – the proposed method based on Mel-frequency cepstral coefficients using neural network classifier.

LPCC-NN – the method based on linear prediction cepstral coefficients using neural network classifier.

MFCC-GMM – the method based on Mel-frequency cepstral coefficients using Gaussian mixture modeling classifier.

TABLE I. The Comparative Characteristics of Accent Recognition Methods

Num. instances	Method	Num. incorrectly recognized instances	Recognition error	Recognition reliability
100	MFCC-NN	9	9%	91%
100	LPCC-NN	21	21%	79%
100	MFCC-GMM	13	13%	87%
500	MFCC-NN	43	8,6%	91,4%
500	LPCC-NN	110	22%	78%
500	MFCC-GMM	58	11,6%	88,4%
1000	MFCC-NN	81	8,1%	91,9%
1000	LPCC-NN	208	20,8%	79,2%
1000	MFCC-GMM	128	12,8%	87,2%

The average values of recognition reliability equal 91,43% for MFCC-NN method, 78,73 % for LPCC-NN method, 87,53 % for MFCC-GMM method.

Thus, the reliability of recognition is increased by

$$\nabla k = (100 - \frac{78,73}{91,43}) = 13,89 \approx 14\% ,$$

as compared to recognition reliability for LPCC-NN method, and by

$$\nabla k = (100 - \frac{87,53}{91,43}) = 4,2 \approx 4\% ,$$

as compared to recognition reliability for MFCC-GMM method.

VIII. Accent Recognition as Part of an ELearning Environment

Accent recognition can be considered as one of the steps to foreign language pronunciation assessment in general. Accent recognition subsystem can be plugged into a known massive open online courses system, such as Moodle[1].

Moodle allows external modules to be used within a course with the help of the so called Activity modules[2]. For the accent recognition subsystem such an activity module accesses an external service written in Python with the help of Django framework (see Fig. 2). The service gets a student utterance of a proposed text, does accent recognition using neural networks described in the paper, and returns a result - the accent classified with the highest reliability score. Interaction with a student is organized in a Web browser with a JavaScript application (see Fig. 3).

IX. Conclusion and Further Work

It is shown that the method based on Mel-frequency cepstral coefficients using neural network can improve the reliability of accent recognition by 14% in comparison with other existing methods.

[1] http://moodle.org

[2] https://docs.moodle.org/dev/Activity_modules

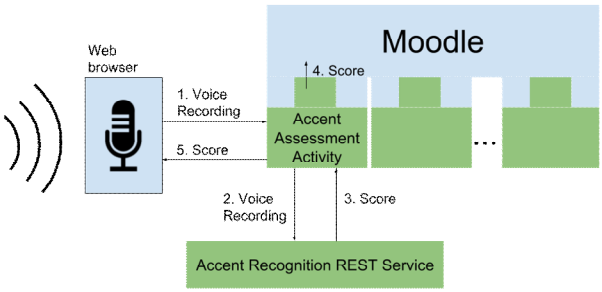

Figure 2. Accent recognition subsystem relations with Moodle.

Figure 3. User interface of accent recognition subsystem: collecting and evaluating utterances.

The speaker accent recognition method is implemented as the external activity module for MOOC Moodle, for the department of English Philology at Zaporizhzhya National University, to serve for students' individual tasks during their study of foreign languages.

Further development of the work is to collect the samples of Slavic languages accents when speaking foreign languages, to create correspondent accent recognition models. It will increase the reliability of recognition of accents and provide more adequate assistance during foreign language learning.

REFERENCES

[1] J. H. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *Proc. ICASSP*, pp. 836-839, 1995.

[2] M. Benzeghiba et al. "Automatic speech recognition and speech variability: A review," *Speech communication* Vol. 49, No.10, pp. 763-786, 2007.

[3] M. Najafian, et al. "Acoustic model selection using limited data for accent robust speech recognition," *Proc. 22nd European IEEE Signal Processing Conference (EUSIPCO)*, pp. 1786–1790, 2014.

[4] C. Teixeira, I. Trancoso and A. Serralheiro, "Accent identification." Proc. 4th Intl. Conf. Spoken Language, Vol. 3, IEEE, pp. 1784–1787, 1996.

[5] M. K. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," *Proc. Int'l Conf. Acoustics Speech and Signal Processing*, IEEE, pp. 4398–4401, 2010.

[6] A. Hanani, M. J. Russell and M. J. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Computer Speech & Language,* 27(1), 2013, pp. 59-74.

[7] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing*, Vol. 4, No.1, p. 31, 1996.

[8] W. M. Campbell et al, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, Vol. 20, No.2, pp. 210-229, 2006.

[9] P. A. Torres-Carrasquillo et al, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *Proc. Interspeech*, pp. 89-92, 2002.

[10] P. A. Torres-Carrasquillo, T. P. Gleason and D. A. Reynolds, "Dialect identification using Gaussian mixture models," *ODYSSEY 2004-The Speaker and Language Recognition Workshop*, 2004.

[11] E. Singer et al, "The MITLL NIST LRE 2011 language recognition system," *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006, 738 p.

[13] P. Angkititrakul and J. H.L. Hansen, "Advances in phone-based modeling for automatic accent classification," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No.2, 2006, pp. 634-646.

[14] H. Tang and A. A. Ghorbani, "Accent classification using support vector machine and hidden Markov model," *Proc. 16th Canadian society for computational studies of intelligence conference on Advances in artificial intelligence*. Springer, Berlin, Heidelberg, pp. 629-631, 2003.

[15] J. Gonzalez-Dominguez et al, "Automatic language identification using long short-term memory recurrent neural networks," *Proc. 15th Annual Conf. of the Int'l Speech Communication Assoc.*, 2014.

[16] J. Gonzalez-Dominguez et al, "Frame-by-frame language identification in short utterances using deep neural networks," *Neural Networks*, Vol. 64, pp. 49-58, 2015.

[17] K.H. Davies, R. Biddulph and S. Balashek, "Automatic Speech Recognition of Spoken Digits," *J. Acoust. Soc. Am., vol.* 24 (6), pp. 637-642, 1952.

[18] G.A. Einicke, *Smoothing, Filtering and Prediction: Estimating the Past, Present and Future*. Rijeka, Croatia, 2012, 286 p.

[19] Zh. Fang, G. Zhang and S. Zhanjiang, "Comparison of Different Implementations of MFCC," *J. Computer Science & Technology*, *vol*. 16(6), pp. 582-589, 2001.

[20] J. Durbin, "The fitting of time series models," *Rev. Inst. Int. Stat.*, Vol. 28, pp. 233–243, 1960.

[21] E.A. Pervushin, "Obzor osnovnykh metodov raspoznavaniya diktorov," *Matematicheskiye struktury i modelirovaniye, vol*. 24, pp. 41-54 (in Russian), 2011.

[22] D. Dowe, Mixture Modeling page. URL: http://users.monash.edu/~dld/mixturemodel.html. Last checked on 26th June, 2017.

[23] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of the IEEE*, Vol. 77(2), pp. 257-286, 1989.

[24] Y. DongSuk, "Robust speech recognition using neural networks and hidden markov models," PhD thesis, 106 p.

[25] N. V Le, D. Panchenko, "Raspoznavaniye rechi na osnove iskusstvennykh neyronnykh setey," *Tekhnicheskiye nauki v Rossii i za rubezhom: materialy Mezhdunar. nauch. konf.*, pp. 8-11, 2011 (in Russian)

[26] Scikits.talkbox. URL: https://scikits.appspot.com/talkbox

[27] Wildcat Corpus of Native- and Foreign-Accented English. URL: http://groups.linguistics.northwestern.edu/speech_comm_group/wildcat/. Last checked on 26th June, 2017.