Published online: 21/08/2014
Published print: 08/2014

doi: 10.5455/aim.2014.22.246-248
ACTA INFORM MED. 2014 AUG 22(4): 246-248
Received: 11 June 2014 • Accepted: 22 July 2014

ORIGINAL PAPER

A HTK-based Method for Detecting Vocal Fold Pathology

Vahid Majidnezhad

Department of Computer Engineering, Shabestar Branch, Islamic Azad University, Shabestar, Iran

Corresponding author: Vahid Majidnezhad. Department of Computer Engineering. Shabestar Branch. Islamic Azad University, Shabestar, Iran. E-mail: vahidmnyahoo.com.

ABSTRACT
Introduction: In recent years a number of methods based on acoustic analysis were developed for vocal fold pathology detection. These methods can be categorized in two categories: a) detection based on the phonemes b) detection based on the continuous speeches. While there are many researches which belong to the first category, there are few efforts for detecting vocal fold pathology based on the continuous speeches (second category). Methods: In this work, a method based on the Hidden Markov model Toolkit (HTK) for detecting vocal fold pathology in the Russian digits is developed which belongs to the second category. It employs a three state HMM for modeling each phoneme. Results: According to the results of the experiments, the proposed method achieves the 90% of detection accuracy. Conclusion: The proposed method is one of the first works for detecting vocal fold pathology based on the Russian digits (from 1 to 10) for Belorussian people. The reported accuracy is rather good and therefore it is recommended to use it as an auxiliary tool in medical centers.
Keywords: vocal fold pathology, Automatic Speech Recognition (ASR), Hidden Markov model Toolkit (HTK), Russian digits.

1. INTRODUCTION

The vocal fold, also known as vocal cord, is a part of sound box. The vocal cords consist of twin infoldings of mucous membrane stretched horizontally across the larynx. It is opened during inhalation, is closed when holding one's breath, and is vibrated during phonation; the folds are controlled by the vagus nerve. The vocal folds are brought near enough together so that air pressure builds up beneath the larynx. The cords are pushed apart through this increased subglottal pressure by the inferior section of each cord leading the superior section. Under the normal conditions, this oscillation pattern will sustain itself. Modulating the flow of air being expelled from the lungs during phonation, the vibration of the vocal folds happens. Rhythmic opening and closing of the vocal folds, leads chopping up of a steady flow of air around the glottal into little puffs of sound waves. An insight about voice production can be achieved by the studies of vocal fold biomechanics; these studies also can provide important information about laryngeal pathology development. A special and great interest is pathology diagnosis for voice production system when there is not any visual evidence for morphological laryngeal abnormalities (1). The process of generating certain sounds through quasi-periodic vibration by the use of vocal folds is called phonation. Dysfunction of the vocal fold in which the phonation process is defected partially or completely is called vocal fold pathology. Due to this disorder, voice quality of a person is altered in such a way that it is thought to be abnormal to the listener. Occurrence of this disorder can be "sudden" or "slow".

There has been a growing interest in detection of vocal fold pathology within international voice communities in recent years. This problem can be categorized in two categories: a) detection of vocal fold pathology based on the phonemes especially vowels b) detection of vocal fold pathology based on the continuous speeches especially digits. There are many works in the first category. For example, in (2) vocal fold pathology was detected using Hidden Markov Model (HMM). In another study (3) vocal fold pathology was detected using Support Vector Machine (SVM). In another study (4) vocal fold pathology was detected using Artificial Neural Network (ANN). In another study (5) vocal fold pathology was detected using Gaussian Mixture Model (GMM). In another study (6) vocal fold pathology was detected by using decision tree. In another study (7) vocal fold pathology was detected by the use of K-Nearest Neighbors (KNN). In another study (8) vocal fold pathology was detected by using Linear Discriminant Analysis (LDA). All of the above mentioned studies used only vowel /a/ as an input. Comparative evaluation between sustained vowel and continuous speech for acoustically discriminating pathological voices was studied in (9). It was found in their experiment that classification of voice pathology was easier for sustained vowel than for continuous speech. But there are few works in the second category such as (10). The aim of this article is to develop a method for Russian digits which belongs to the second group.

Nowadays, Automatic Speech Recognition (ASR)

has grown rapidly. One of the well-known ASR tools is HTK. In the current study a conventional ASR system, HTK, was used for detecting vocal fold pathology in speaking Russian digits. MFCC (Mel Frequency Cepstral Coefficients) and GMM (Gaussian Mixture Model)/HMM (Hidden Markov Model) were used as features and classifier, respectively. The authors of this article believe that this is the first such work that tries to detect vocal fold pathology for Russian digits.

2. MATERIALS AND METHODS

2.1. HTK

HTK is a toolkit for building Hidden Markov Models (HMMs) 11. HMMs can be used to model any time series and the core of HTK is similarly general-purpose. However, HTK is primarily designed for building HMM-based speech processing tools, in particular recognizers. Thus, much of the infrastructure support in HTK is dedicated to this task.

In fact, Hidden Markov Model (HMM) is a statistical model. It consists of a finite number of hidden states which poses some observations with their respective occurrence probabilities. The states are not visible, but the observations are visible. So, the sequence of observations can be used for extracting some information about the sequence of states. HMMs are used successfully for modeling the stochastic process and consequently in processing of biomedical signals. It can be used for classification tasks especially in bioinformatics and signal processing. If for each class and its respective data set, a model is constructed and trained. Then, these models can be used for classifying new data.

In the HTK, there are two major processing stages involved. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions. Secondly, unknown utterances are transcribed using the HTK recognition tools.

HTK was originally developed at the Machine Intelligence Laboratory (formerly known as the Speech Vision and Robotics Group) of the Cambridge University Engineering Department (CUED) where it has been used to build CUED's large vocabulary speech recognition systems. Using HTK usually involves the following steps:

Step 0 – Preparing the initial files: For training the HMMs, a dictionary should be made to define the valid words and their pronunciation for the recognition.

Step 1–Feature Extraction: In this phase the raw speech data (signal waveforms) are transformed into sequences of feature vectors.

Step 2–Training the HMMs: The parameters of the HMMs are trained using the iterative EM-algorithm and the obtained feature vectors from the former step.

Step 3–Recognizing Test Data: In the HTK there is a general-purpose Viterbi word recognizer. It matches speech signals against a network of HMMs and returns a transcription for each speech signal.

Step 4–Calculating recognition accuracy: The real class labels and the recognized class labels are compared to calculate the recognition hit rates.

2.2. Mel Frequency Cepstral Coefficients (MFCCs)

In recent years, the MFCCs features are used to characterize speech signals. They can be estimated by the use of a parametric approach derived from the linear prediction coefficients (LPC), or by the use of non-parametric discrete fast Fourier transform (FFT) that encodes usually more information than the LPC method. The speech signal is windowed with a Hamming window in the time domain and by the use of the FFT converted into the frequency domain which gives the magnitude of the FFT. Then the FFT data is converted into the filter bank outputs and the cosine transform is found to reduce dimensionality. The filter bank is made by using the 13 linearly-spaced filters (133.33Hz between center frequencies,) followed by the 27 log-spaced filters (separated by a factor of 1.0711703 in frequency). Each filter is made by the combination of the amplitude of FFT bin.

2.3. Data set

The data set was created by specialists from the Belorusian Republican Center of Speech, Voice and Hearing Pathologies. In recording, the utterers read specially picked up text (also including Russian digits) within several minutes. About 5 hours of records of speech of healthy people and about 10 hours of records of patients with pathologies have been recorded. All of the samples are the wave files in the PCM format and in a mono mode and the sample rate of 44100 Hertz and the bit-depth of 16 bit. We have chosen speech samples of 50 subjects including 25 normal subjects and 25 pathological subjects. A total of 50 speakers uttered the ten Russian digits (from 1 to 10). All speakers were native Belarusian. Two sets of data were created: one for training the ASR system with 30 records and the other for testing with 20 records.

3. EXPERIMENTS AND RESULTS

The experiment in this work was conducted on a connected phoneme task constituting isolated Russian digits. Each phoneme was modeled by a three state HMM. Observation probability density functions were modeled using GMM. All the training and recognition experiments were implemented with the HTK package.

The parameters of the system were: 25 milliseconds Hamming window with a frame period of 10 milliseconds, and the pre-emphasis coefficient was 0.97. As features, delta and acceleration coefficients are to be computed and appended to the static MFCC coefficients.

For displaying and comparing the results, four indicators (TP, FN, TN and FP) have been used. True

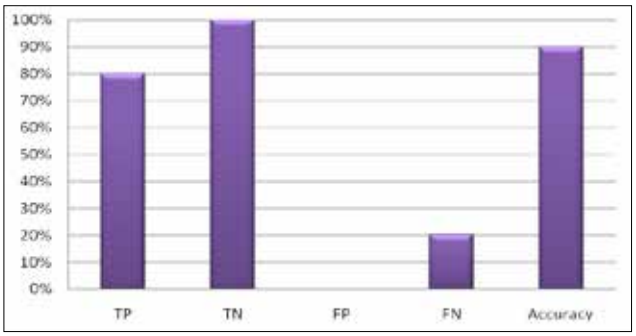

Figure 1. *The results of experiment in our research*

positive rate (TP), also called sensitivity, is the ratio between pathological files correctly classified and the total number of pathological voices. False negative rate (FN) is the ratio between pathological files wrongly classified and the total number of pathological files. True negative rate (TN), sometimes called specificity, is the ratio between normal files correctly classified and the total number of normal files. False positive rate (FP) is the ratio between normal files wrongly classified and the total number of normal files. The final accuracy of the system is the ratio between all the hits obtained by the system and the total number of samples. The recognition performances of the proposed method are shown in the Figure 1. As it can be seen in the Figure 1, the obtained TP, TN, FP, FN rates are 80%, 100%, 0%, 20% respectively. And alsothe final obtained accuracy is 90%.

4. CONCLUSION

In this article, a HTK-based method for detecting vocal fold pathology was proposed that it uses continuous speech, Russian digits, as the input of system. In the proposed method, Mel-Frequency-Cepstral-Coefficients (MFCC) with the delta and acceleration coefficients was used as the observation of the system. Observation probability density functions were modeled by means of GMM. A three state HMM was used for modeling the phonemes. The HTK package was used for the training and recognition processes. Russian digits ASR performance based on the HTK was evaluated. According to the results of experiments, the recognition accuracy of 90% was achieved.

Although it may be possible to try to build a complete multiclass classification system so that detection of different type of pathological speech will be possible.

Acknowledgements

This work was supported by the speech laboratory of the United Institute of Informatics Problems of NASB in Belarus. The authors wish to thank the Belorusian Republican Center of Speech, Voice and Hearing Pathologies by its support in the speech database.

CONFLICT OF INTEREST: NONE DECLARED.

REFERENCES

1. Cveticanin L. Review on Mathematical and Mechanical Models of the Vocal Cord. Journal of Applied Mathematics. 2012; 1: 1-18.
2. Majidnezhad V, Kheidorov I. A HMM-Based Method for Vocal Fold Pathology Diagnosis. IJCSI International Journal of Computer Science Issues. 2012; 9(2): 135-138.
3. Majidnezhad V, Kheidorov I. The SVM-Based Feature Reduction in Vocal Fold Pathology Diagnosis. International Journal of Future Generation Communication and Networking. 2013; 6(1): 45-56.
4. Majidnezhad V, Kheidorov I. An ANN-based Method for Detecting Vocal Fold Pathology. International Journal of Computer Applications. 2013; 62(7): 1-4.
5. Majidnezhad V, Kheidorov I. A Novel GMM-Based Feature Reduction for Vocal Fold Pathology Diagnosis. Research Journal of Applied Sciences, Engineering and Technology. 2013; 5(6): 2245-2254.
6. Lee JY, Jeong S, Hahn M. Pathological Voice Detection Using Efficient Combination of Heterogeneous Features. IEICE TRANS. INF. & SYST. 2008; E91-D(2): 367-370.
7. Hariharan M, Paulraj P, Yaacob S. Identification of Vocal Fold Pathology based on Mel Frequency Band Energy Coefficients and Singular Value Decomposition. The International IEEE Conference on Signal and Image Processing Applications, (ISCIPA 2009), Kuala Lumpur, Malaysia. Nov. 2009: pp 514-517.
8. Kaleem MF, Ghoraani B, Guergachi A, Krishnan S. Telephone-quality Pathological Speech Classification using Empirical Mode Decomposition. The 33rd Annual International Conference of the IEEE EMBS, Boston, Massachusetts USA. September 2011: pp 7095-7098.
9. Parsa V, Jamieson DG. Acoustic Discrimination of Pathological Voice: Sustained Vowels versus Continuous Speech. Journal of Speech, Language and Hearing Research. 2001; 44: 327-339.
10. Muhammad G, Mesallam T, Malki KH, Farahat M, Alsulaiman M, Bukhari M. Formant analysis in dysphonic patients and automatic Arabic digit speech recognition. BioMedical Engineering OnLine. 2011; 10(41): 1-12.
11. Young S. The HTK Book (for HTK Version. 3.4). Cambridge University Engineering Department, 2006.