GRINREY

Advanced Research in Computer Engineering

Sandip A. Kale
Editor

Research Transcripts in Computer, Electrical and Electronics Engineering

4 Development Software for Preprocessing Voice Signals

Niyozmatova N[a,*] Mamatov N[a], Nurimov P[a], Samijonov A[a] and Samijonov B[a]

[a]Tashkent University of Information Technologies named after Al-Kharezmi, Tashkent, Uzbekistan

*Corresponding author: n_nilufar@mail.ru

ABSTRACT

At present, the most important task of modern science is the creation for a person of natural means of communication with a computer, where speech input of information is carried out in the most convenient way for the user. Speech recognition is one of the challenges. As practice shows, the quality of recognition depends on the properties of the preprocessing system. To improve the quality of recognition, it is necessary to develop effective and high-speed methods and algorithms for signal preprocessing. This article proposes a new approach and algorithm for extracting features of speech signals. Based on the proposed algorithm, the identification problem is solved. In addition, the chapter presents a description of the software module for each stage of the preliminary processing of speech signals. This software is a voice-based identity tool.

Keywords: *speech signal, filter, MFCC, PLP, LPCC*

Advanced Research in Computer Engineering | ISBN: 978-81-948951-2-1
Book Series: Research Transcripts in Computer, Electrical and Electronics Engineering | Vol: 02

1. INTRODUCTION

Speech is a sequence of sounds. Sound is a superposition of sound waves (vibrations) with different frequencies. From the point of view of physics, a wave is characterized by amplitude and frequency. Speech signal processing is a field of science where noise suppression, amplification, filtering, information extraction, information separation, compression, restoration and coding are carried out. It has become widespread in all areas of speech technology.

2. METHODOLOGY AND SOLUTION OF THE PROBLEM

The software for pre-processing speech signals consists of several stages; each stage includes several modules (Fig. 1).

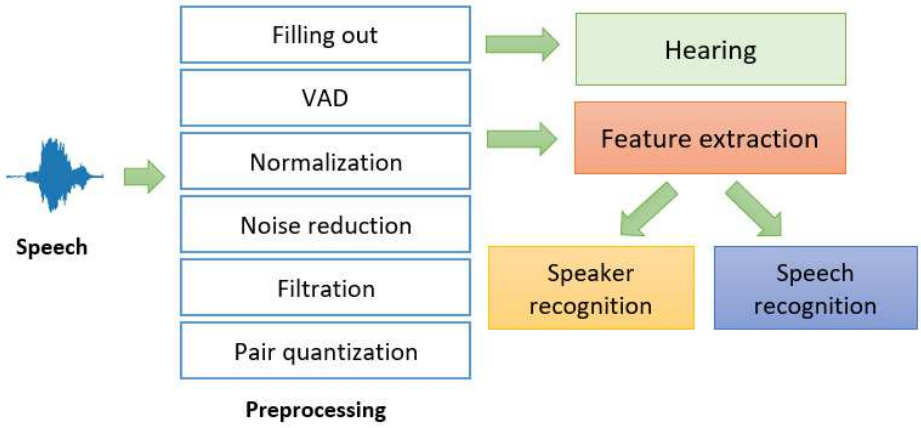

Fig. 1. The structure of the software for preprocessing speech signals

2.1. The speech signals

The first stage of the system under development is “Speech Signals”. At this stage, voice signals received from a file or through a microphone. At the first stage, speech signals received from a file or through a microphone.

2.2. Preprocessing of speech signals

At this stage, the signal are initially processed by the following modules,

2.2.1 Filling out

Gaps in speech signals are filled.

2.2.2 Voice Activation Detection-VAD

This module is used to find speech in audio. VAD is used to reliably detect speech in audio and even in background noise. In completely pure audio recordings, even elementary energy detection may be sufficient for speech recognition; but unfortunately, in the wild there are not always completely pure signals, so VAD must be noise resistant. VAD consists of the steps:

- Signal division into frames
- Formation of features for each frame
- Classifier training in active and quiet frames
- Classification of invisible frames as speech or silence

An example on the created software using the function of detecting voice activity for a speech signal is shown in Fig. 2 and Fig.3 [1].

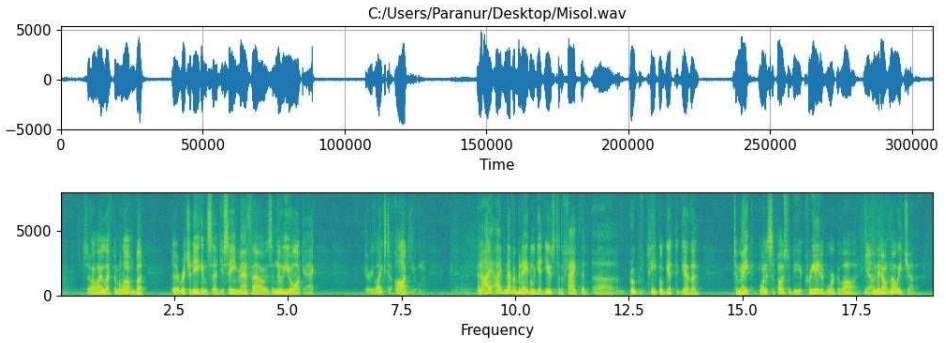

Fig. 2. Speech signal before applying VAD

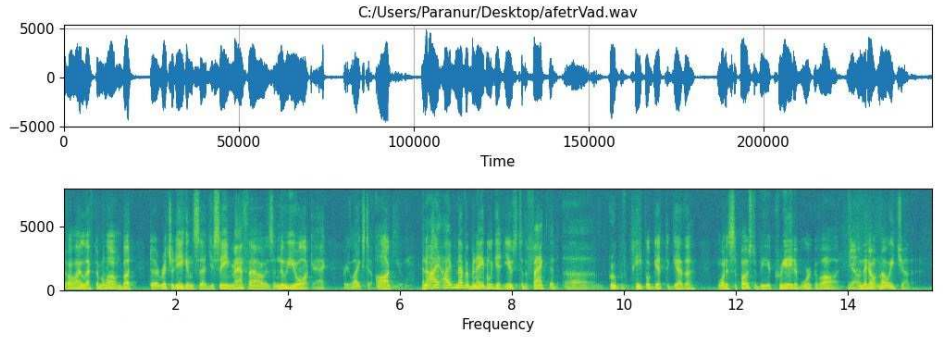

Fig. 3. Speech signal after applying VAD

2.2.3 Normalization

EBU R128 volume normalization consists of dynamic and linear modes. Support is provided in the form of single-pass (live broadcasts, files) and two-pass (files) modes. Typically, this algorithm targets IL, LRA, and the largest true peak. If the normalization mode is not linear, then the audio stream will be

sampled to 192 kHz to accurately determine the true peaks. To explicitly set the sample rate of the output, we need to use the -ar option or a sample filter [2]. The filter has the following variable parameters (Fig. 4):

- I, i - setting the built-in volume range. Here, the default values are -24.0, and the interval [-70.0; -5.0].
- LRA, lra - setting the target volume range. Here the default values are 7.0 and the interval is [1.0; 20.0].
- TP, tp - setting the maximum true peak. Here the default values are -2.0 and the interval is [-9.0; 0.0].

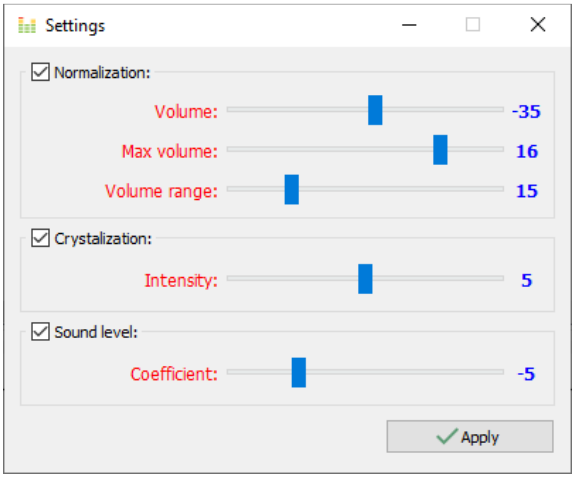

Fig. 4. Normalization parameters

2.2.4 Noise reduction

In [2], noise suppression of audio samples with FFTs is described in detail. In experimental studies, the parameters were taken as follows:

- nr is noise reduction (dB), the default value is 12 (dB), and the allowable range is nr from 0.01 to 97.
- nf is the minimum noise level in dB, where the allowable range is between -80 and -20. The default is -50 dB.
- nt is the type of noise that has the parameters: wn is white noise, Vn is vinyl noise, sn is shellac noise and cn is the user noise defined in option bn. nt parameter defaults to white noise.
- bn is the user noise range which is each of the 15 bands, where the bands are separated by '' or '|'.
- rf is the residual level (dB), where the permissible range of rf is between -80 and -20. The default is -38 (dB).
- tn - noise tracking, which takes the values 1 (true) or 0 (false). The default is false. With this power on, the noise level is automatically adjusted.
- tr - track balances, enable or disable. The default is 0 (Fig. 5).

2.2.5 Filtration

Filtering is one of the main stages of processing time or spatial series of measurements. Currently, there are many filtering methods, for example, median filtering, polynomial approximation, cosine filtering, Fourier transform and wavelet transform, etc. (Fig. 6).

The Savitsky-Golay filter is digital

Based on the Savitsky-Golay filter, a set of digital data points is carried out for data smoothing, which allows increasing the data accuracy without signal distortion. Precision is maximized in the process, that is, as in convolution, by selecting successive subsets of adjacent points using a low degree algebraic polynomial based on the least squares method. If the points are located at the same distance, then the analytical solution to the least squares equations can be found in a single set of "convolution coefficients" that can be applied to all subsets of the data to obtain smoothed estimates. The signal (or derivatives of the smoothed signal) at the center point of each subset. This method, based on established mathematical procedures, was popularized by Abraham Savitsky and Marcel J.E. (Fig. 7, Fig 8.)

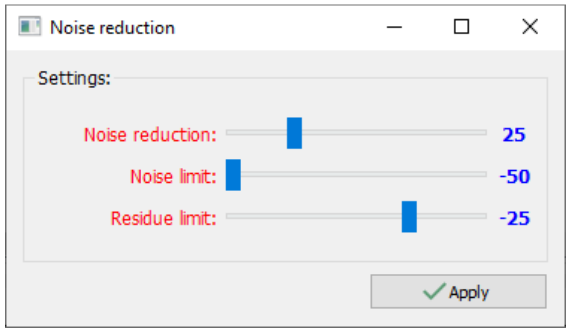

Fig. 5. Window of Noise Reduction

The created software uses the savgol_filter [3] function in the Python programming language. Filtering parameters are presented in the form below.

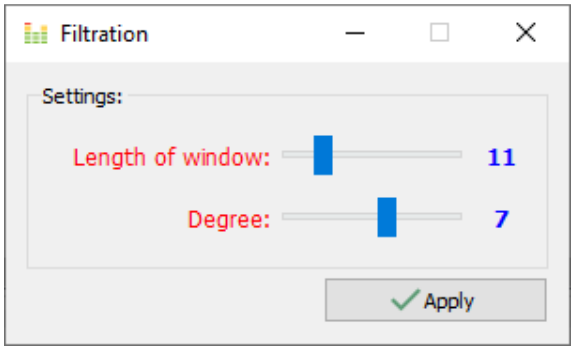

Fig. 6. Window of filtration

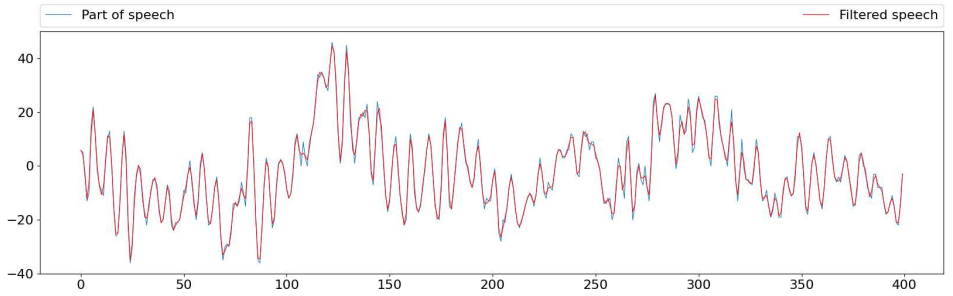

Fig. 7. An example of applying the filter of Savitsky-Golay. Win_len = 5, poly_order = 3

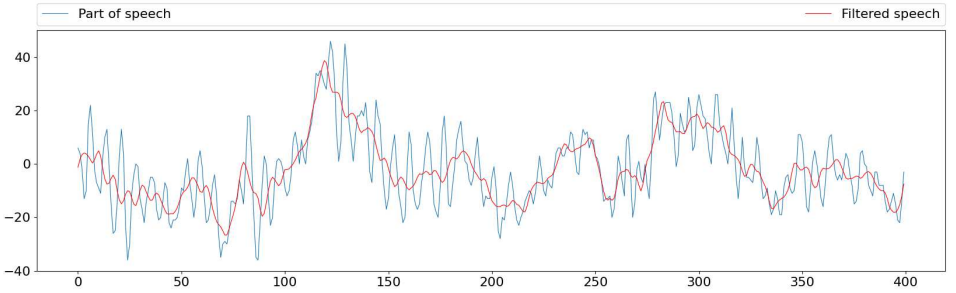

Fig. 8. An example of applying the filter of Savitsky-Golay. Win_len = 15, poly_order = 3.

2.2.6 Pair quantization

The pair quantization, carried out by reducing the number of proposed features, is implemented as follows. Features are distinguished from the speech signal based on the following formula (pattern),

$$s_i = \frac{1}{L} \sum_{j=i \cdot L}^{(i+1) \cdot L} x_j \qquad (1)$$

where x_j is a vector consisting of symbols of a given speech signal, L -step.

2.3. Hearing

At this stage, the pre-processed speech signal can be saved to a file in audio format.

2.4. Feature extraction

At this stage, the features of speech signals are formed as: MFCC, LPC, PLP and medium quantum.

The linear prediction cepstral coefficient method, the perceptual linear prediction coefficient method and robust PLP (PLP-RASTA), the cepstral

coefficient coefficient method on the chalk scale (MFCC) are the most powerful methods based on cepstral signal analysis.

LPCC is a linear prediction cepstral coefficient method. Based on the calculation of the coefficients of the autoregressive model for each frame of the audio signal. After obtaining all the model parameters, cepstral coefficients are calculated based on the recursive function.

PLP is a perceptual linear prediction coefficient method. The method differs from the LPCC method in that it takes into account the characteristics of the perception of various frequencies by a person - before calculating the parameters of the autoregressive model, the signal undergoes a certain pre-processing. The calculated instantaneous Fourier spectrum is converted into a spectrum on the barque scale, after which the operation of convolution of the masking curves of the critical bands with the obtained spectrum is performed to obtain the frequency masking effect. Next, the volume curve and cepstral processing are approximated.

The advantage of the PLP method compared to LPCC is that it allows you to suppress information related to the individual characteristics of the speaker by choosing the appropriate model order. However, this method is more sensitive to the pitch frequency.

2.4.1 Mel-coefficient coefficients (MFCC)

MFCC - based on human auditory perception, and obtained on a scale of twisted frequencies. To calculate the MFCC, a speech window is first created to divide the speech signal into frames. To get the same amplitude for all formats, high frequency formants are reduced in amplitude, compared to low frequency formants, high frequencies are emphasized. After creating a window for calculating the power of the spectrum for each frame, a fast Fourier transform is performed. After application (FFT) based on the filter base using the melting scale, it is processed by the power spectrum. To calculate the MFCCs after transforming the power spectrum to the logarithmic domain, DCT is performed in the speech signal. The following formula is used to calculate Mel for an arbitrary frequency [4, 5]:

$$mel(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

where $mel(f)$ -Mel frequency, a f - frequency.

MFCC is calculated using the following formula:

$$\hat{C}_n = \sum_{n=1}^{l} \left(\log \hat{S}_l\right) \cos\left[n\left(l - \frac{1}{2}\right)\frac{\pi}{l}\right] \tag{3}$$

where $l-$ number of Cepstral melting factors, $\hat{S}_l$ - output filter set, $\hat{C}_n$ are the last coefficients of MFCC.

Based on MFCC, the low-frequency region is determined more efficiently than the high-frequency region, and it can calculate formants belonging to the low-frequency range, and also describes the resonances of the vocal tract.

MFCC is universal and recognized as an interface procedure for typical identification applications [4]. In addition, MFCC is ideal when the source characteristics are stable and consistent across sounds [6]. MFCC is also able to capture information from sampled signals in which the frequency is not more than 5 kHz. Such signals cover most of the energy of sounds generated by humans.

In speech recognition, MFCC is widely used [7]. There are formants higher than 1 kHz and are not taken into account due to the large distance in the high frequency range between filters [4]. In the presence of MFCC background noise, the signs are not accurate and cannot be generalized [5].

2.4.2 Perceptual Linear Prediction (PLP)

Based on the PLP method, critical bands are combined, the intensity is compressed into loudness and a preliminary emphasis on equal loudness when extracting relevant information from speech.
PLP is based on a non-linear scale and was first used in speech recognition problems to eliminate speaker-dependent functions [8]. PLP represents a smoothed aligned and compressed, similar to human hearing, short-term spectrum, transforms it as MFCC.
Based on the PLP approach, some important hearing characteristics can be expressed, and the subsequent auditory speech spectrum is approximated using an autoregressive pan-polar model [9]. With PLP, the minimum high frequency resolution is obtained, which means the auditory filterbank approach, but produces orthogonal results like cepstral analysis. For spectral smoothing, PLP uses linear prediction, hence the name - perceptual linear prediction [10]. The combination of spectral and linear predictive analysis is PLP.
To calculate the characteristics of PLP, speech is highlighted as a window (Hamming window), the Fast Fourier Transform (FFT) is calculated and squared. This gives energy-spectral estimates. Then a trapezoidal filter is applied at a predetermined interval, usually 1 cortex.
This filter integrates overlapping responses and also effectively squeezes the highest frequencies into a narrow band. As a result of integration, high frequencies are effectively compressed into a narrow band.
Then, on the scale of distorted frequencies of the cortex, convolution is performed in the symmetric frequency domain, which allows smoothing the spectrum, where low frequencies mask with high frequencies.
Then the symmetrical frequency domain is convolved according to the scale of distorted frequencies of the cortex, which, by smoothing the spectrum, allows low frequencies to mask high frequencies. Subsequently, the spectrum is pre-emphasized to approximate the uneven sensitivity of human hearing at different

frequencies. Compression of the spectral amplitude decreases the change in the amplitude of spectral resonances. IDCT is performed to obtain autocorrelation coefficients. Spectral smoothing is carried out on the basis of solving autoregressive equations. With the help of autoregression coefficients, the values of cepstral parameters are obtained [10].

To calculate the frequency of the cortex, the following formula is used:

$$bark(s) = \frac{26.81s}{1960+s} - 0.53 \tag{4}$$

where $bark(s)$ - core frequency, and s - frequency in Hz.,

The identification results based on PLP are much better than LPC [10]. It justifies that PLP effectively suppresses speaker-specific information. In addition, it has improved independent recognition characteristics and is resistant to noise, changes in channels and microphones.

Based on PLP, autoregressive noise components are accurately restored. In addition, PLP is more sensitive to any changes in the frequency of formants.

Linear Prediction Cepstral Coefficients (LPCC). LPCC is the cepstral coefficient, which is obtained from the calculated envelope of the LPC spectrum. The LPCC coefficients are illustrations of the Fourier transform of the logarithmic spectrum of quantities [11, 12] LPC.

Cepstral analysis ideally symbolizes speech signals and characteristics with a limited size of functions and therefore they are usually used in the field of speech processing [12].

Rosenberg and Sambur noted that adjacent predictor coefficients have a high correlation, and with less correlated characteristics the representations will be more effective, therefore LPCC is of this kind.

If the signals have a minimum phase, then the LPC is easily converted to LPCC [13].

In speech processing based on LPCC, they are likewise computed as LPC [12].

LPCC are calculated based on the following formula [13]:

$$C_m = a_m + \sum_{k=1}^{m-1} \left[\frac{k}{m}\right] c_k \, a_{m-k} \tag{5}$$

here a_m - linear prediction and C_m - cepstral coefficient.

LPCCs have a lower error rate than LPC functions [12] and have a slight noise vulnerability [11]. Higher order cepstral coefficients are mathematically limited, resulting in an extremely large amount of variation when going from lower order cepstral coefficients to higher order cepstral coefficients. Likewise, LPCC estimates are highly sensitive to quantization noise.

In high frequency speech signal, cepstral analysis gives little separability between source and filter in the field of quantity. The lowest order cepstral coefficients are sensitive to spectral tilt, and the highest order coefficients are noise sensitive.

The result of this step, i.e. the processed speech signal is transmitted to the identification step. If the identification result is lower than expected, proceed to stage 3 and the process continues.

2.5. Speaker recognition and speech recognition

At this stage, the processed speech signal is used to identify a person by voice. If the result is lower than the specified result, go to stage 3, and the process continues.

Vector quantization (VQ). VQ [15] is an efficient data compression method. It is effectively and successfully applied in various systems such as vector quantization coding and recognition.

Codebooks are created on the basis of the LBG algorithm [14, 15]. In [16, 17], the following steps of the LBG algorithm are given:

i. Development of the 1st vector codebook; which is the centroid of all training vectors.

ii. Doubling the size of the codebook is done according to the rule by dividing each current codebook c_m:

$$c_m^+ = c_m(1+\varepsilon) \tag{6}$$

$$c_m^- = c_m(1-\varepsilon) \tag{7}$$

where m ranges from 1 to the current codebook size, ε - split parameter.

iii. Finding centroids for a shared codebook

iv. Repeat stages 2 and 3 until a codebook of size M is developed.

Euclidean distance

Based on the Euclidean distance, the similarity or difference between two spoken words is calculated, arising after quantizing these words in its codebook. The new word is compared by measuring the Euclidean distance between the feature vector of the new word and the model (codebook) of known words in the base. The word with the smallest distance is chosen according to the following formula:

$$d(x,y) = \sqrt{\sum_{i=1}^{M}(x_i - y_i)^2} \tag{8}$$

where, x_i is the *i*-th vector input features, y_i is the *i*-th vector features in the codebook, d is the distance between x_i and y_i.

Based on the proposed structure (Fig. 1), software for personal identification by voice (Fig. 10) and speech recognition has been developed. The general view of the developed software is shown in Fig. 9.

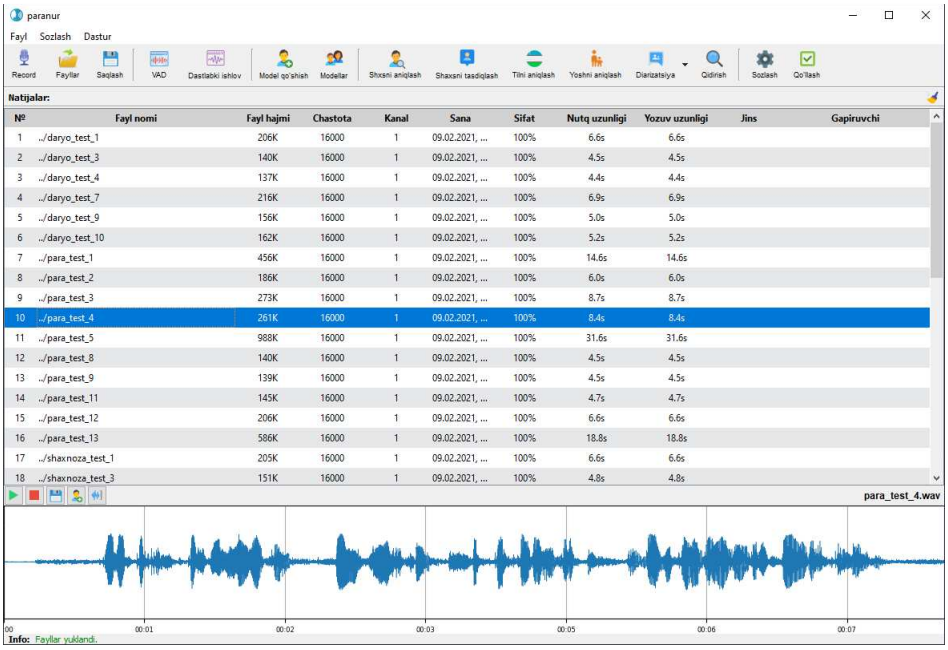

Fig. 9. Window software preprocessing speech signals

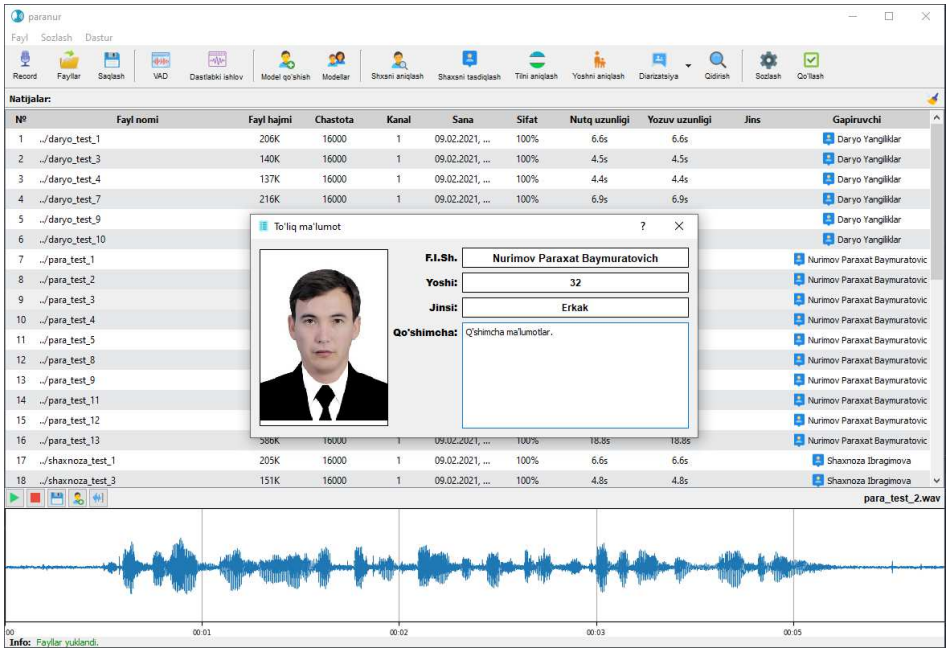

Fig. 10. Identification module

3. RESULTS

The base of speech given in [18, 19, 20] consists of 100 people, of which 74 are men and 26 are women. The database contains 10 speech files for each person, 6 of which were used for training and the remaining 4 for testing. In this case, the duration of each file is 2-5 seconds.

The experimental results were obtained on the basis of the above-mentioned speech base using the main algorithm and the proposed algorithms.

Table 1. The results of the comparison algorithms

Algorithm	Data Size (MB)	Results (%)	Time of training (sec.)	Time spent on identification (sec.)
Basic	285,3	99,6	45,3	69,3
The proposed	31,5	98,4	6,7	13,1

4. DISCUSSION

Experimental results show that the accuracy of the proposed algorithm is less than 2% than that of the existing algorithm. Experimental studies have shown that the famous algorithms do not meet the requirements of real time. The algorithm proposed in the work, by reducing the number of features used in personal identification by voice, made it possible to reduce the recognition time by up to 5 times. And also the size of the file used for recognition is significantly reduced. This ensures a reduction in the volume of the speech database and an increase in the speed of data transmission over the channels. This justifies the use of the program based on the proposed algorithm in recognition systems.

5. CONCLUSION

The chapter proposes a pair quantization algorithm. The number of signs obtained using this algorithm is at least two times less than the number of signs obtained using existing algorithms. This speeds up identification. Based on the proposed algorithm, experimental studies were carried out using an example of solving a practical problem. As a result of the work done, a software structure for preprocessing speech signals for a speech recognition and identification system is proposed. It is planned to develop a system for automatic speech recognition based on the attributes of a pair quantization algorithm.

REFERENCES

[1] https:github.com/wiseman/py-webrtcvad/

[2] https:ffmpeg.org/ffmpeg-filters.html#loudnorm

[3] https:docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.signal.

savgol_filter.html

[4] Chakroborty S, Roy A, Saha G. Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification. In: IEEE International Conference on Industrial Technology, 2006. ICIT 2006. pp. 387-390

[5] Hasan MR, Jamil M, Rabbani G, Rahman MGRMS. Speaker Identification Using Mel Frequency cepstral coefficients. In: 3rd International Conference on Electrical & Computer Engineering, 2004. ICECE 2004. pp. 28-30

[6] Chu S, Narayanan S, Kuo CC. Environmental sound recognition using MP-based features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE; 2008. pp. 1-4

[7] Ravikumar KM, Reddy BA, Rajagopal R, Nagaraj HC. Automatic detection of syllable repetition in read speech for objective assessment of stuttered Disfluencies. In: Proceedings ofWorld Academy Science, Engineering and Technology. 2008. pp. 270-273

[8] Ravikumar KM, Rajagopal R, Nagaraj HC. An approach for objective assessment of stuttered speech using MFCC features. ICGST International Journal on Digital Signal Processing, DSP. 2009;9(1):19-24

[9] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America. 1990;87(4):1738-1752

[10] Kumar P, Chandra M. Speaker identification using Gaussian mixture models. MIT International Journal of Electronics and Communication Engineering. 2011;1(1):27-30

[11] El Choubassi MM, El Khoury HE, Alagha CEJ, Skaf JA, Al-Alaoui MA. Arabic speech recognition using recurrent neural networks. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No.03EX795). Ieee; 2003. pp. 543-547. DOI: 10.1109/ISSPIT.2003.1341178

[12] Wu QZ, Jou IC, Lee SY. On-line signature verification using LPC cepstrum and neural networks. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. 1997; 27(1):148-153

[13] Holambe R, Deshpande M. Advances in Non-Linear Modeling for Speech Processing. Berlin, Heidelberg: Springer Science & Business Media; 2012

[14] Linde Y., Buzo A. and Gray R.M. An algorithm for vector quantizer design. IEEE Trans. Communication, 1980. vol. COM-28, no. 1, pp. 84-95.

[15] A. Gersho, R. M. Gray Vector Quantization and Signal Compression. Kluwer Academic Publishers, Boston, MA, 1991.

[16] H. B. Kekre, Vaishali Kulkarni, Speaker Identification by using Vector

Quantization. International Journal of Engineering Science and Technology 2010 Vol. 2(5), pp 1325- 1331.

[17] M. Narzillo, S. Abdurashid, N. Parakhat, and N. Nilufar, “Automatic speaker identification by voice based on vector quantization method,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 10, pp. 2443–2445, 2019.

[18] B. Wiedecke, M. Narzillo, M. Payazov, and S. Abdurashid, “Acoustic signal analysis and identification,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 10, pp. 2440–2442, 2019.

[19] M. Narzillo, S. Abdurashid, N. Parakhat, and N. Nilufar, “Karakalpak speech recognition with CMU sphinx,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 10, pp. 2446–2448, 2019.

[20] https:www.dropbox.com/s/87v8jxxu9tvbkns/development_set.zip?dl=0

Cite this article

Niyozmatova N, Mamatov N, Nurimov P, Samijonov A and Samijonov B, Development Software for Preprocessing Voice Signals, In: Sandip A. Kale editor, Advanced Research in Computer Engineering, Pune: Grinrey Publications, 2021, pp. 53-66