Algorithms and architectures of speech recognition systems

N Mekebayev[1], O Mamyrbayev[2], M Turdalyuly[3], D Oralbekova[4], M Tasbolatov[5]

[1,2]Institute of Information and Computational Technologies, Kazakhstan

[1,2,5]al-Farabi Kazakh National University, Almaty, Kazakhstan

[3,4]Satbayev University, Almaty, Kazakhstan

[5]Kazakh national women's teacher training university, Almaty, Kazakhstan

nurbapa@mail.ru

ABSTRACT

Digital processing of speech signal and the voice recognition algorithm is very important for fast and accurate automatic scoring of the recognition technology. A voice is a signal of infinite information. The direct analysis and synthesis of a complex speech signal is due to the fact that the information is contained in the signal.

Speech is the most natural way of communicating people. The task of speech recognition is to convert speech into a sequence of words using a computer program.

This article presents an algorithm of extracting MFCC for speech recognition. The MFCC algorithm reduces the processing power by 53% compared to the conventional algorithm. Automatic speech recognition using Matlab.

Keywords

speech recognition, audio signal, neural network, MFCC, ASR

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

Automatic speech recognition is a dynamically developing area in the field of artificial intelligence.

Speech recognition is the process of recognizing words spoken by a person based on an automatic speech signal. The length of a word can be different at different frequencies, and the same length depends on the same words, different parts of words are different from the apparent rate of difference in the environment. You need to align the time to get the distance between the speeches (represented as vector sequences). The comparison of the concept of dynamic alignment by spectral sequence of words has been used to solve the problems.

The problem of speech recognition is a current problem today.

Most modern methods used to solve it require large computational resources, the amount of which is often limited. The impossibility of wide application of many algorithms today, for example, in mobile devices makes researchers look for more effective methods.

This article describes the algorithm and analysis of speech recognition methods, the identification of the shortcomings of each of them. Development of the program for speech recognition and experiment.

The speech recognition module includes two main digital signal processes: function extraction and function matching. The first one processes the word spoken by the user and generates its functions. The spoken speech is first converted into a digital domain, and the digital sampled speech is processed to extract functions using the MFCC approach, which evaluates the vocal track filter. In section I pre-processing of the speech signal is discussed, in section II we consider the selection of speech characteristics using the MFCC algorithm , the architectures of the automatic speech recognition system are considered in section III.

Preprocessing The Speech Signal

When recording speech, the amplitude of the sound signal is influenced by a number of factors: the volume of the speaker's voice, its distance from the microphone, etc. These factors lead to a large variability in the volume of the speech signal [1]. This phenomenon is especially noticeable when using a heterogeneous sound recording equipment To eliminate the volume spread, the amplitude normalization procedure is applied. Using this technique, the amplitude of a signal lies in boundaries$[\frac{\Delta}{2},-\frac{\Delta}{2}]$ (Fig.1) $\widehat{p}[n]$ is performed by the following formula:

$$\widehat{p}[n] = \frac{\Delta}{\max\limits_{m}|s[m]|} \cdot p[n] \qquad (1)$$

where Δ –normalization band width , symmetric about the abscissa axis (for example, in Fig. 1. Δ= 1).

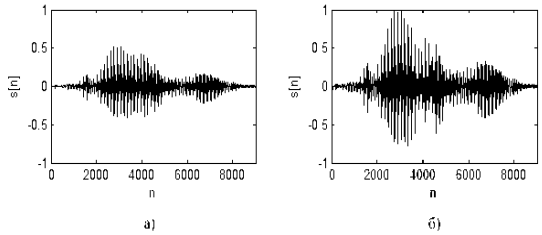

Figure 1. The digitized speech signal before (a) and after (b) normalization .

To estimate the volume variability, we will consider a set of Q examples of uttering one or more words. Let's find the average value of the volume M_q [2] for the i-th example with length of N samples and average M_Q for Q examples:

$$\mathrm{M(q)} = \sum_{n=1}^{N}|p[n]|, q = 1, \dots, Q; \qquad (2)$$

$$\mathrm{M_Q} = \frac{1}{Q}\sum_{q=1}^{Q} M(q) \qquad (3)$$

After that, we calculate the relative deviation (q) of the volume of each example from the mean:

$$D_q = |1 - \frac{M(q)}{M_Q}| \quad (4)$$

It can be seen from the formulas (2), (3) that both the absolute value of the count and the number of these samples influence the result in the example, therefore it is necessary to estimate the variability of the volume of the set of examples of one class, he length of which is approximately the same, and the variation in the loudness of the base as a whole. Figure 2.a)-b) shows D ( ) for Q = 100 recorded examples of the word "three" before and after normalization, respectively. The volume spread was 28.5% for the original examples and 14.3% for the normalized ones. Graphs 2.b)-d) show (q) for speech database from Q = 2000 examples of different utterances. The volume spread was 24.8% for the original database and 23.11% - normalized.

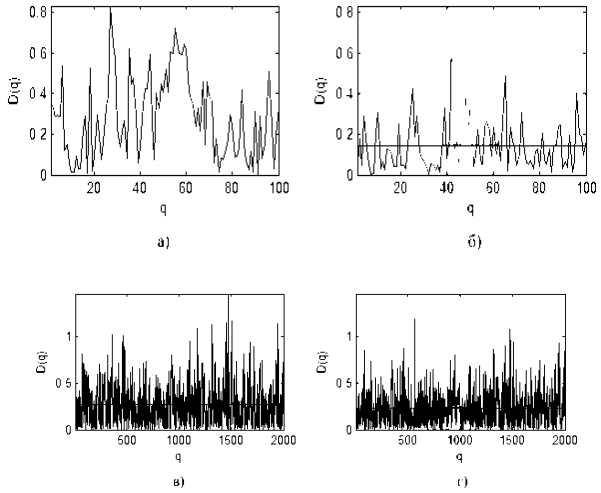

Figure 2. The deviation of the example energy from the average energy value for (a, b) original examples and normalized (b, d).

As can be seen from the study, the use of normalization always allows to reduce the volume spread for different utterances These results were obtained for a speech database collected under the same conditions on the only available equipment, so in general, normalization played a minor role. However, normalization is necessary in the operation of a real system, when the speech signal is received under different conditions.

Feature Extraction Of Speech Using Mfcc Algorithm

Thus, a signal sounds at the input of our system. The sound is divided into frames - sections of 25 ms with overlapping frames equal to 10 ms.

For processing the audio signal should be converted into a signal spectrum, or in the form of logarithmic spectrum, with subsequent scaling, since this corresponds to the human perception of sound (Mel-scale). Then the signal is represented in the form of MFCC (Mel-frequency cepstral coefficients) by applying a discrete cosine transform. MFCC is usually a vector of thirteen real numbers, it represents the energy of the signal spectrum. This method takes into account the wave nature of the signal, the mel-scale allocates the most significant frequencies perceived by the person, and the number of MFCC coefficients can be set by any number, which allows to compress the frame and reduce the amount of information processed [3].

We will consider the algorithm of MFCC-conversion of the received audio signal. The received audio signal is discretized:

$x[n], 0 \le n < N.$ (5)

We represent it as the Fourier transform:

$$X_Q[k]=\sum_{n=0}^{N-1} x[n] e^{\frac{-2\pi i}{N}kn}, 0 \le k < N. \quad (6)$$

We calculate the filter comb using the window:

$$H_m = \begin{cases} 0 & k < f[m-1]; \\ \frac{(k-f[m-1])}{(f[m]-f[m-1])} & f[m-1] \le k < f[m]; \\ \frac{(f[m+1]-k)}{(f[m+1]-f[m])} & f[m] \le k < f[m+1]; \\ 0 & k > f[m+1], \end{cases} \quad (7)$$

where f [m] is equal to

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B((f_1) + m\frac{B(f_h)-B(f_1)}{M+1}\right) \quad (8)$$

We represent our frequencies in the form of a Mel-scale in (b):

$$B^{-1}(b) = 700\left(exp\left(\frac{b}{1125}\right)^{-1}\right) \quad (9)$$

where the energy of the windows will be

$$S[m] = ln\left(\sum_{k=0}^{N-1} |X_Q[k]|^2 H_m[k]\right), 0 \le m < M \quad (10)$$

We obtain the MFCC coefficients:

$$c[n] = \sum_{N=0}^{M-1} S[m] cos\left(\pi n\left(m+\frac{1}{2}\right)/M\right),$$

$0 \le n < M$ (11)

Let our frame be represented as a discrete value vector according to the formula (9).

We calculate the spectrum of the signal:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2\pi kn}{N}}, 0 \le k < N \quad (12)$$

$$H[k] = 0.54 - 0.46 \cdot cos\left(\frac{2\pi k}{N-1}\right) \quad (13)$$

$X[k]=X[k]\cdot H[k], 0 \le k < N$ (14)

On the OX axis, the frequency is plotted in Hertz, the magnitude-on the axis OY , so as not to bind to complex quantities (Figure 3):

Mel representation shows the significance of individual sound frequencies for a person, depends on the specific frequencies of sound, and on the volume, and on the timbre of a person. Mel-scale is calculated as follows (forward and reverse conversion):

$$M = 1127 \cdot log\left(1 + \frac{F}{700}\right) \quad (15)$$

$$F = 700 \cdot (e^{M \cdot 1127} - 1) \quad (16)$$

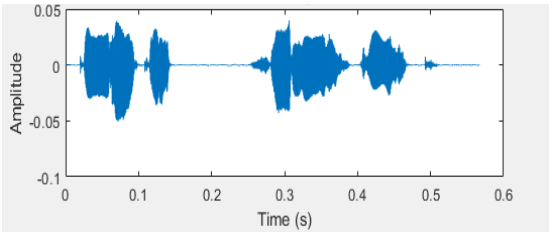

Figure 3. Representation of the original signal as the Fourier transformation

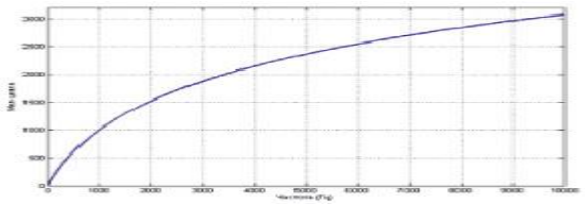

Figure 4. The graph of Mel-scale versus frequency

The graph of the Mel-scale versus frequency is shown in Fig. 4.

These computing units are the most common in speech recognition systems, because they correspond to the peculiarities of human perception of sound.

We will consider an example: a frame with a length of 256 samples, the sound frequency of 16 kHz are given. Let human speech be concentrated in the frequency range from 300 Hz to 8 kHz. The most frequently used number of Mel-coefficients is ten, and we will use it.

At first it is necessary to calculate a comb of filters to present the spectrum in the mel-scale format. The Mel filter is a triangular window that sums the energy at its frequency range and calculates the mel coefficients. Since we know the number of coefficients, we can build a set of ten filters (Figure 5).

In the low frequency range (the frequencies that we are most interested in), the number of windows is larger, which provides high resolution. This can significantly improve the quality of recognition.

In order to find the signal energy, we multiply the signal spectrum vector and the window function, as a result of which we obtain the coefficients vector. If they are squared and represented as a logarithm and obtained from them by cepstral coefficients, then we obtain the desired mel-coefficients.Cepstral coefficients can be obtained using the Fourier transformation, and the discrete cosine transformation [4].

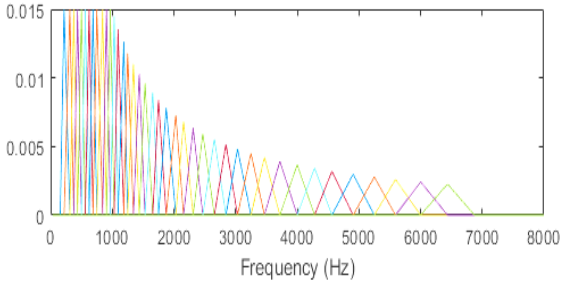

Figure 5. Mel-frequency cepstral coefficients

The frequency range is from 300 Hz to 8 kHz. On the mel-scale, this range corresponds from 401.25 to 2834.99. Now we build twelve control points for the construction of ten triangular filters (Mel-scale and the scale is in Hertz):

$m[i] = [400.25; 622.50; 1066.00; 1286.25; 1507.50; 1727.74; 1939.98; 2161.25; 2392.49; 2612.73; 2835.98]$ (17)

$h[i] = [310; 517.35; 782.90; 1105.97; 1499.04; 1973.32; 2554.33; 3261.62; 4122.63; 5160.75; 6445.70; 801$ (18)

As we have already said, the frame length is 256 samples of the signal, the frequency is 16 kHz (it is plotted along the OX axis). We will put the calculated scale on the signal spectrum.

$f(i)=floor((frameSize+1)*h(i)/sampleRate)$ (16)

which corresponds

$f(i)$=4;8;12;17;23;31;40;52;66;82;103;128 (17)

By control points we will build filters:

$$H_m = \begin{cases} 0 & k < f[m-1]; \\ \frac{(k-f[m-1])}{(f[m]-f[m-1])} & f[m-1] \le k < f[m]; \\ \frac{(f[m+1]-k)}{(f[m+1]-f[m])} & f[m] \le k < f[m+1]; \\ 0 & k > f[m+1], \end{cases} \quad (19)$$

The filter is multiplied with the spectrum:

$$S[m] = \log\left(\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right),$$

$0 \le m < M$ (20)

Mel-filters are applied to the energy of the spectrum, then the resulting values are taken the logarithm. The discrete cosine transformation is applied to obtain the cepstral coefficients, it compresses the obtained results, increases the contribution of the first coefficientsand reduces the contribution of the latter.

$$C[l] = \sum_{m=0}^{M-1} S[m] cos\left(\pi l \left(m + \frac{1}{2}\right)/M\right),$$

$0 \le l < M$ (21)

It turns out that we have 12 coefficients (Fig. 5). s a result, a small finite set of values (for example, twelve coefficients in our case) allows us to replace the use of a huge numeric array of signal samples, either the signal spectrum or the signal periodogram. Each word of finite length corresponds to a set of Mel-frequency cepstral coefficients. Then it is necessary to find the closest model for a certain set of Mel-frequency cepstral coefficients. For this we are looking for the Euclidean distance between the vector of chalk-frequency cepstral coefficients and the vector of the model under study. The required model is the one with the smallest calculated distance.

The set of MFCC coefficients for the same word may differ, for example, if the word is pronounced by two different people, or the pronunciation speed is different. For these purposes, the algorithm of dynamic time transformation is used. It calculates the optimal time deformation between the compared time sequences [5].

	-2	**10**	**-10**	**15**	**-13**	**20**	**-5**	**14**	**2**
3	5	12	25	36	53	70	78	89	90
-13	16	28	15	43	37	70	78	105	104
14	32	20	39	16	43	43	62	62	74
-7	37	37	23	38	21	49	45	66	71
9	48	38	42	29	44	33	47	50	57
-2	48	50	46	46	40	55	36	52	55

Figure 6. Results of calculations

Experimental Part

I have written the program in the Matlab environment and conducted an experimental study of the algorithm for recognizing speech signals, the task was to collect data on word recognition: "Torequl ..." Each word was uttered three times by three people. The noise threshold was set, although the noise was insignificant, but it still could affect the

results. The results of the statistical data are given in the tables.

Table 1 Data on the word "Torequl ...."

	Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5
Man 1	+	-	+	+	-
Man 2	-	+	+	+	+
Man 3	-	+	+	-	+
Man 4	+	+	-	-	+
Man 5	-	+	-	-	-
Man 6	+	+	-	+	+
Man 7	-	+	+	+	+

In total, it turns out that the recognition percentage of the word "Torequl ..." is about 75%.

Algorithm Of The Program

At the beginning of the work, the main program window is displayed. Then an audio message is sent to the microphone dynamic speaker which is responsible for the input module of the voice signal. Then, the user selects the program mode on the main window. If the reference creation mode is selected, for which the module of the reference databasea creation is responsible, the program processes and stores the input signal from the microphone and displays the spectrum on the screen. If the recognition mode is selected, the program processes the results and compares them with the pre-recorded reference in the database, stores the input signal and proceeds to its recognition by calculating the first and second finite difference of the total phase function, i.e. we determine the number of sounds in this word, as can be seen from the previous simulation. We determine the start and end of the word by selecting the envelope. The recognition result is displayed. Figure 11 shows a schematic view of the program. The scheme is shown below.

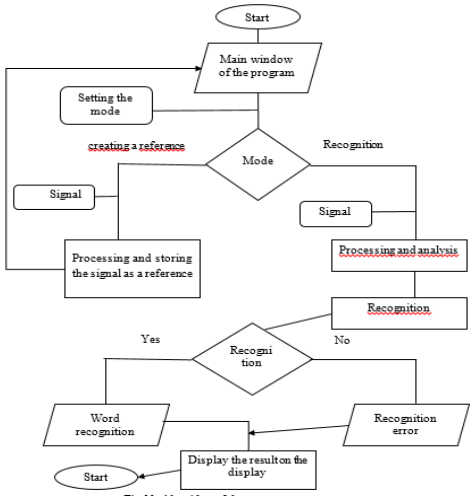

Figure 11. Algorithm of the program

Conclusion

It was noted that the MFCC for each individual user is unique. Certain variations were observed due to differences in the locality of the recording area. These MFCC are then compared, that is, the MFCC pattern and the real-time input are compared for each user. In programming, the Euclidean distance is used to compare the pattern and the input in real time. Thus, the MFCC algorithm is used for speech recognition.

The approach discussed was tested on the voices of different people. The database was created for the voice of 7 different persons, corresponding to the numbers from 0 to 9 and some control words. The MFCC coefficients corresponding to these training voice samples were stored together with the calculated weights. MFCC is calculated for the test sample. The article has been written on the basis of the project

Acknowledgements.

IRN-AR05131207 "Development of the technology of multilingual automatic speech recognition using deep neural networks".

References

[1] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) Techniques" Issue 3, March 2010.

[2] Watcher, M. D., Matton, M., Demuynck, K., Wambacq, P., Cools, R., "Template Based Continuous Speech Recognition", IEEE Transaction on Audio, Speech, & Language Processing, 2007.

[3] Gupta, R., and Sivakumar, G., "Speech Recognition for Hindi Language", IIT BOMBAY, 2006.

[4] Ingle V., Proakis J. Digital Signal Processing Using Matlab V4 – Boston: ITP, 1997.

[5] Rabiner, L. Juang, B. H., Yegnanarayana, B., "Fundamentals of Speech Recognition", Pearson Publishers, 2010.

[6] Barsky A.B., Neural networks: recognition, management, decision-making.

[7] Cheong Soo Yee and abdul Manan ahmad, Malay Language Text Independent Speaker Verification using NN-MLP

classifier with MFCC, 2008 international Conference on Electronic Design.

[8] Wu Junqin, Yu Junjun, "An Improved Arithmetic of MFCC in Speech Recognition System," IEEE Transaction on Audio Speech processing, and Language, pp.719-722, 2011.

[9] Gurpreet Kaur, Harjeet Kaur, "Multi Lingual Speaker Identification on Foreign Languages Using Artificial Neural Network with Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 2013.

[10] L. Rabiner and G. Juang, "Fundamentals of Speech Recognition," Prentice-Hall, 1993.

[11] H. Hasegawa, M. Inazumi, "Speech Recognition by Dynamic Recurrent Neural Networks," Proceedings of 1993 International Joint Conference on Neural Networks.

[12] Furui, S., 50 years of progress in speech and speaker recognition. SPECOM 2005, Patras, 2005: pp. 1-9.

[13] Mamyrbayev O, Toleu A, Tolegen G, Mekebayev N. "Neural Architectures for Gender Detection and Speaker Identification" Journal Cogent Engineering. ISSN: 2331-1916. Volume 7, 2020 - Issue 1

[14] Bagher BabaAli, Waldemar Wojcik, Orken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev. Speech Recognizer-Based Non-Uniform Spectral Compression for Robust MFCC Feature Extraction // Przeglad Elektrotechniczny. ISSN: 0033-2097 – 2018. - № 6 (94). – P. 90-93. (Scopus) (Clarivate Analytics)

[15] Orken Mamyrbayev, Mussa Turdalyuly, Nurbapa Mekebayev. Automatic Recognition of Kazakh Speech Using Deep Neural Networks // Lecture Notes in Computer Science. 11432 LNAI, c. 465-474.

[16] Orken Mamyrbayev, Turdalyuly, Nurzhamal Oshanova,Tolga Ihsan Medeni, Aigerim Yessentay. Voice Identification Using Classification Algorithms // We are IntechOpen, the world's leading publisher ofOpen Access books Built by scientists, for scientists. June 25, 2019. London.

[17] Kalimoldayev, M., Mamyrbayev, O., Mekebayev, N., Kydyrbekova, A. Algorithms for detection gender using neural networks // International Journal of Circuits, Systems and Signal Processing. 2020

[18] Mamyrbayev, O., Alimhan, K., Zhumazhanov, B., Turdalykyzy, T., Gusmanova, F. End-to-End Speech Recognition in Agglutinative Languages Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioin formatics), 2020

[19] Aizat, K., Mohamed, O., Orken, M., Ainur, A., Zhumazhanov, B. Identification and authentication of user voice using DNN features and i-vector // Cogent Engineering. 2020