A New Method Towards Speech Files Local Features Investigation

Rustam Latypov, Evgeni Stolov
Kazan Federal University
Kazan, 420008, Russia
Roustam.Latypov@kpfu.ru

Abstract—There are a few reasons for the recent increased interest in studying the local features of speech files. It is stated that many of the essential properties of the speaker's language used may appear in the form of a speech signal. The traditional instruments — short Fourier transform, wavelet transform, Hadamard transforms, autocorrelation, and the like can detect not all peculiar properties of the language. In this paper, we propose a new approach to such characteristics exploration. The original signal is approximated by a new one, which values are taken from a finite set. We then construct a new sequence of fixed-size vectors based on these approximations. Studying the distribution of generated vectors provides a new way of describing the local attributes of speech files. Finally, the developed technique is applied to the problem of the automated distinguishing of two known languages used in speech files. For this, a simple neural net is constructed.

Keywords—Language distinguishing; Local properties; Neural net; Speech files

I. INTRODUCTION

Voice recognition software can reduce keyboard usage, which speeds up queries. If the request is addressed to a person, there are no problems with recognizing the language used. Consider a case where a country has two or more official languages. A citizen forms a speech request using his preferred language to a state body. Generally, the request cannot be answered immediately, and the authority wants a printed version of the message. Various services can solve this problem, but before converting the query to a print form, the system must determine one of the possible languages used in the request. There are many publications dedicated to the mentioned challenge. In [1], a review of the earlier implemented methods is presented. A set of phonemes corresponding to each of the considered languages is taken. The information system compares the current phoneme from the speech file with the ones installed in the set. There are visible drawbacks to this approach. Since the templates of the chosen phonemes depend on the speakers, so many templates from different speakers should be presented. An option of the described procedure is comparing signals spectra in place of comparing the signals in the time domain [2]. More recently, neural network-based language identification systems (LID) approaches are increasingly popular and reported to offer superior performance compared to traditional LID techniques [3]. Lately, end-to-end multilingual systems based on the sequence-to-sequence architecture are proposed [4]. A part of language parameters can be used for a unified architecture with a shared vocabulary among many languages. Compelling systems have recently developed, which can recognize speech narrated in one of the known languages [5], [6]. Those systems need significant resources, and only a restricted number of public authorities can afford such a system. There is a chance to use some unique features in files that afford to determine used language without vast calculation if we deal with two languages that have different origins. That is the case we investigate in this paper: distinguish Tatar and Russian speech files. It is known that languages differ in voice onset time (VOT) [7], [8]. The works, exploiting this peculiarity in files, suggest various methods for localization of consonants in speech. One is measuring the distance from the consonant’s end to the beginning of vocalization [9]. The principal obstacle is obtaining the moment where vowel begins. It can be performed via analysis of the spectrum of the signal inside the window, moving along file [10]. Another approach to the problem is the supposition that some sounds produced by native speakers have some peculiarities. Revealing such peculiarities, one can distinguish the languages. It was shown in [11] that the waveform, corresponding to consonants, can be used as a mark for solving the problem. A parametric curve approximates the wave, and the parameters providing the best approximation determine the language with high reliability. The form of the wave also defines some other features of the speech. [12] shows that the signal samples in the extreme value area determine the instant frequency of the signal. Hence these set some unique features of the speaker. In this paper, we continue the investigation of the local signal properties. A kind of correlation must be presented among neighboring samples in speech. The standard procedures can not reveal that correlation since a signal from the consonant part of the file can not be viewed as a stationary one. To overcome that obstacle, we use the following procedure. We move a window lengthways signal. We approximate the signal inside the window with the values from a particular set. Then we change the window contents with a vector with items in this set. After that, we train a neural net to restore the original signal via the vector. This net is utilized as a filter for investigation of the local file properties. That determines the language used by the speaker.

978-1-7281-5973-7/20/$31.00 ©2020 IEEE

62nd International Symposium ELMAR-2020, 14-15 September 2020, Zadar, Croatia
II. APPROXIMATION OF SIGNAL

Let a discrete source signal be a sequence $\text{Sign} = \langle a_0, a_1, \ldots, a_{n-1} \rangle$. We are going to replace the original signal with a stepwise sequence $\text{Appr} = \langle b_0, b_1, \ldots, b_{N-1} \rangle$. In this article, we consider the case when a stepwise sequence has five levels. First, we define four thresholds: $Th_0 < Th_1 < Th_2 < Th_3$. The rule for matching the original signal $a$ to level $b$ is defined by (1).

$$b = \begin{cases} 
2, & a \geq Th_3 \\
1, & a \geq Th_2 & a < Th_3 \\
0, & a \geq Th_1 & a < Th_2 \\
-1, & a \geq Th_0 & a < Th_1 \\
-2, & \text{otherwise} 
\end{cases}$$

(1)

To compare the approximation $\text{Appr}$ with the original signal $\text{Sign}$, we need to choose a metric. This allows optimizing the choice of $\text{Appr}$. As usual, we use the standard SNR to this end (2):

$$\text{SNR}(\text{Sign}, \text{Appr}) = 10 \log_{10} \left( \frac{\sigma^2(\text{Sign})}{\sigma^2(\text{Sign} - \text{Appr})} \right),$$

(2)

where $\sigma^2$ is variance and both the arrays are supposed to have equal standard deviations. If we go the standard way, then we need to implement the optimization procedure with four variables to get the optimal values for the thresholds in metrics (2). Instead, for optimization, we leverage suboptimal procedures that lead to acceptable results. The source signal $\text{Sign} = \text{Sign}_+ + \text{Sign}_-$, where the first term contains non-negative and the second term – non-positive items. Each of the terms is approximated separately. We set a relation between the thresholds, that reduce the optimization procedure to the tabulation of a single parameter with a step. We suppose that $Th_2 > 0, Th_1 < 0$ and set $Th_3 = 2 \cdot Th_2$ and $Th_0 = 2 \cdot Th_1$. The suboptimal value for $Th_2$ is calculated by the function $\text{getOptThresh}$ where $In = \text{Sign}_+$. The $Th_1$ value is obtained using the same procedure, where $In = -\text{Sign}_-$ and $Th_1$ equals the inverse value returned by the procedure. The realization of $\text{getOptThresh}$ is presented in Procedure 1. Here the function $\text{getAppr}$ converts the source signal according (3)

$$b = \begin{cases} 
2, & a \geq 2 \cdot \text{Thresh} \\
1, & a \geq \text{Thresh} & a < 2 \cdot \text{Thresh} \\
0, & \text{otherwise} 
\end{cases}$$

(3)

into a new file with items in $Dset$ defined by (4)

$$Dset = \{-2, -1, 0, 1, 2\}.$$  

(4)

Let $\text{Appr}$ with items in $Dset$ be the approximation of the source signal $\text{Sign}$. It is calculated by the formula (5)

$$\text{Appr} = \text{getAppr}(\text{Sign}_+, \text{Thresh}_2) - \text{getAppr}(-\text{Sign}_-, -\text{Thresh}_1).$$

(5)

An example of such approximation is shown in Fig. 1.

III. VECTOR PRESENTATION OF STEPWISE SIGNAL

Our goal is to study the dependency of neighbor samples in speech files. Since the assumption that the signal is stationary is not fulfilled, the standard correlation is not suitable for this. Our approach to the problem is as follows. We create $\text{Appr}$ and choose a window of length $LW$, which slides along $\text{Appr}$. Any position of the window gives a vector of length $LW$ containing items inside the window. Note that any element of the resulting vectors is in $Dset$ (4). We represent the vector $\text{Vec} = \langle c_0, c_1, \ldots, c_{LW-1} \rangle$ by an integer

$$I_{\text{Vec}} = \sum_{k=0}^{LW-1} c_k \cdot 5^k.$$  

(6)

Thus, we have established a one-to-one correspondence between the constructed vectors and the set of integers in the interval $[-(5^{LW}-1)/2, (5^{LW}-1)/2]$. Now, we can implement standard tools to study digital sequences. Let us create a normalized histogram (the size of each bin is divided into

---

Procedure 1: $\text{getOptThresh}$(In, StartThresh, Step)

1: $\text{SNR, BestThresh} \leftarrow -100, 0 \triangleright$ Save best value of SNR and Threshold
2: $\text{NIn} \leftarrow \text{In} \triangleright$ Normalize by the standard deviation
3: $\text{Thresh} \leftarrow \text{StartThresh}$
4: while $\text{Thresh} < \text{Max}(\text{NIn})$ do
5: $\text{Fl} \leftarrow \text{getAppr}(\text{NIn}, \text{Thresh})$
6: $\text{NFL} \leftarrow \text{Fl} \triangleright$ Normalize by the standard deviation
7: $\text{Val} \leftarrow \text{SNR}(\text{NIn}, \text{NFL})$ \triangleright Implement (2)
8: if $\text{Val} > \text{SNR}$ then
9: $\text{SNR, BestThresh} \leftarrow \text{Val, Thresh}$
10: end if
11: $\text{Thresh} \leftarrow \text{Thresh} + \text{Step}$
12: end while
13: return $\text{BestThresh}$

---

Figure 1. Approximation of signal by stepwise sequence. II – the source signal, I2 – the approximation
Figure 2. Normalized histograms for various lengths of window

(a) \( LW = 5 \)  
(b) \( LW = 7 \)  
(c) \( LW = 9 \)  
(d) \( LW = 7 \)

Figure 3. Normalized numbers of the appearance of each vector in the file. Length of window equals 7

(a) Speech file  
(b) Noise file

Figure 4. Lengths of intervals with constant components. L1 – Russian speaker, L2 – Tatar speaker

IV. DISTRIBUTION OF VECTORS AND RECOGNITION OF THE LANGUAGE USED IN RECORDS

The main goal of the research in this section is a test of the following hypothesis. In essence, a record’s language can be recognized based on the distribution of vectors if it is known in advance that only one of two languages is used. Our primary tool is the distribution of the vectors among various frequencies, which is shown in Fig. 6. In our case, the database has a limited size, since it is created by direct recording during television news. It contains 31 files with a Russian male voice, 36 files with a Tatar male voice, 26 files with a Russian female voice, and 24 files with a Tatar female voice. All files are
II. We suppose that these results are acceptable since the input average vector. The gained results are shown in Tables I and are utilized for testing recognition of the language via the rows in any matrix are used for training. The rest of the rows

Here the summation is carried over all vectors with positive

The function

While summation, we limited ourselves

The function

Usage of all vectors leads to an average vector being close

TABLE I. CORRECT RECOGNIZED LANGUAGES. RUSSIAN AND TATAR MALE SPEAKERS

|        | Russian, total of 16 | Tatar, total of 18 |
|--------|----------------------|--------------------|
| 16     | 12                   |                    |
| 11     | 12                   |                    |
| 8      | 11                   |                    |
| 16     | 10                   |                    |
| 16     | 12                   |                    |

TABLE II. CORRECT RECOGNIZED LANGUAGES. RUSSIAN AND TATAR FEMALE SPEAKERS

|        | Russian, total of 13 | Tatar, total of 12 |
|--------|----------------------|--------------------|
| 13     | 9                    |                    |
| 13     | 12                   |                    |
| 11     | 12                   |                    |
| 12     | 9                    |                    |
| 12     | 10                   |                    |

ACKNOWLEDGMENT

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University. The Russian Science Foundation grant No19-18-00202 also supports this work.

REFERENCES

[1] M. A. Zissman, K. M. Berking, “Automatic language identification,” Speech Comm., vol. 35, pp. 115–124, 2001.
[2] S. G. Koolagudi, D. Rastogi, R.S. Rao, “Spoken language identification using spectral features,” Communications in Computer and Information Science, vol.306, pp.496-507, 2012.
[3] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, “Automatic language identification using deep neural networks,” IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 5337 - 5341, 2014.
[4] S. Watanabe, T. Hori, and J. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 265 - 271, 2017.
[5] H. Li, B. Ma, and K. Lee, “Spoken Language Recognition: From Fundamentals to Practice,” Proc. of the IEEE, vol. 101, no 5, pp. 1136 - 1159, 2013.
[6] J. Li, L. Deng, R.Haeb-Umbach, and Y. Gong, Robust Automatic Speech Recognition: A Bridge to Practical Applications, Elsevier, 2015.
[7] A.Ali, J. Spiegel, and P.Mueller,“Acoustic–phonetic features for the automatic classification of stop consonants,” IEEE Trans. Speech and Audio Proc., vol. 9, no. 8, pp. 833-841, 2001.
[8] A.L. Francis, V. Ciocca, J. M. C. Yu, “Accuracy and variability of acoustic measures of voicing onset,” J. Acoustical Society of America, vol.113, pp. 1025–1032, 2003.
[9] K. Chenausky, C. Nelson, and H. Tager-Flusberg, “Vocalization Rate and Consonant Production in Toddlers at High and Low Risk for Autism,” Journal of Speech, Language, and Hearing Research, vol. 60, no 4, pp. 865–876., 2017.
[10] R. Latypov, R. Nigmatullin, and E. Stolov, “Automatic spoken language identification by digital signal processing methods. Tatar and Russian languages,” Communications in Computer and Information Science, Springer, ICIST 2017, pp.539-549, 2017.
[11] R. Latypov, R. Nigmatullin, and E. Stolov, “Classification of Speech Files by Waveforms,” Lobachevskii Journal of Mathematics, vol. 36, no. 4, pp. 488-494, 2015.
[12] R. Latypov, E. Stolov, “Instantaneous frequency and detection of dynamics in speech,” In 5th International Symposium ELMAR-2017, Zadar, Croatia, pp.141–144, 2017.