World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:10, 2014

Performance Evaluation of Acoustic-Spectrographic Voice Identification Method in Native and Non-Native Speech

E. Krasnova, E. Bulgakova, V. Shchemelinin

***Abstract*—**The paper deals with acoustic-spectrographic voice identification method in terms of its performance in non-native language speech. Performance evaluation is conducted by comparing the result of the analysis of recordings containing native language speech with recordings that contain foreign language speech. Our research is based on Tajik and Russian speech of Tajik native speakers due to the character of the criminal situation with drug trafficking. We propose a pilot experiment that represents a primary attempt enter the field.

***Keywords*—**Speaker identification, acoustic-spectrographic method, non-native speech.

I. Introduction

NOWADAYS, Russian forensic phoneticians often need to analyze recordings that contain speech by non-native language speakers. In Russia, due to the character of the criminal situation with drug trafficking, one of the most interesting in this respect is Russian speech of Tajik native speakers [1]. One of the main methods used for forensics comparison of recordings is the acoustic-spectrographic method [2]. Non-native speech has some specific phonetic characteristics. This is noted as one of the conclusions in the paper of Bhattacharjee and Sarmah [3] that discusses English speech of Indians. We can rightfully suppose that these characteristics can be the reason for the different variability of phonetic features of speech units, which influences the numeric values of some parameters, which in turn affects the use of the acoustic-spectrographic speaker identification method in case of non-native speech. Performance evaluation of the acoustic-spectrographic speaker identification method in this case and exploration of phonetic features of vowels in non-native speech is important for correct interpretation of the result obtained after the analysis by experts during forensic examination. In our research, we evaluated the performance of the method and explored vowel formant values as a phonetic feature of speech in a non-native language.

E. Krasnova is with the Speech Technology Center, Krasutskogo-4, St.Petersburg, 196084, Russia (e-mail: krasnova@speechpro.com).

E. Bulgakova is with the Department of Speech Information Systems, National Research University of Information Technologies, Mechanics and Optics (ITMO University), Saint-Petersburg, Russia (e-mail: bulgakova@speechpro.com).

V. Shchemelinin is with the Department of Speech Information Systems, National Research University of Information Technologies, Mechanics and Optics (ITMO University), Saint-Petersburg, Russia (e-mail: shchemelinin@speechpro.com).

In this paper, we describe the results of the research in which we focused on the error rate of the acoustic-spectrographic identification method in comparing non-native speech recordings with the ones containing native speech. The pairs of recordings we examined were selected by automatic speaker identification methods as the most difficult.

We would like to point out that this experiment is a pilot and represents a primary attempt to test our hypothesis. Only one expert was involved, and a modest selection of material was used. If our proposal is confirmed, we will proceed with more insightful and substantial research in this field.

II. Acoustic-Spectrographic Voice Identification Method

In this part of the paper, we give an overview of the acoustic-spectrographic voice identification method [2]. This is the implementation of the acoustic-spectrographic method that was used in our research.

A. Selection of Fragments

To compare the phonograms using the acoustic-spectrographic voice identification method the expert needs to detect phonetically similar realizations of the same vowels. This means that the pronounced allophone and its phonetic context must be the same. These realizations can be searched for aurally or instrumentally. According to the method, there should be several comparable realizations of each vowel type. After the utterances are selected the expert enters directly upon the analysis.

B. Analysis

For each pair of vowels the expert analyzes and compares principally vowel formant frequencies (usually 3-5 formants). The expert can also pay attention to the fundamental frequency and the duration of the surrounding phones. When comparing vowel formant frequencies, the same part of formant tracks should be analyzed: the stationary part, the transitional starting part or final transitional part. To obtain the formant frequency values an expert can use a sonogram, but to make the most precise measurements of the frequency values it is advisable to use the average spectrum of the analyzed fragment [4]. The expert's identification decision (whether the two voices belong to the same speaker) is based on the obtained result about the formant frequency values. 7 % is taken as the threshold of within-speaker variation: if the difference between the values of the same formant is more

than 7 % and such difference is systematic, the speakers are probably different.

Thus, the core of the described voice identification method is checking the similarity of a speaker's way of producing the same vowels in the same phonetic contexts.

III. Material and Experiment

A. Material

The experiment was carried out on pairs of recordings selected by an automatic speaker identification method. A large speech database was processed by the automatic identification method and 40 pairs compared with maximum error were selected for the research.

An i-vector based text-independent identification system was used as the automatic speaker identification method [5]. NIST SRE 2012 competitions [6] demonstrated that today systems based on representing a speaker voice model in the total variability space are predominant. The method uses a Gaussian mixture distribution for modeling a speaker's voice and then it reduces the Gaussian mixture distribution to i-vectors in low dimensional space of total variability.

A special preprocessing module was used in the identification system. This module contains an energy-based speech detector and a clipped signal detector [7]. Mel-frequency cepstrum coefficients (MFCC) with their 1st and 2nd order derivatives (39 elements in total) were used as speech features. The length of each speech frame for MFCC calculation was 22 ms with an 11 ms shift. Hamming window was used for Gibbs effect compensation. Compensation of the distortion channel effect was done by cepstral mean subtraction (CMS).

At the stage of voice modeling, we used a gender-independent universal background model (UBM) with a 512-component GMM, obtained by standard ML-training on the telephone part of the NIST SRE 1998-2010 datasets [8, 9]. To accelerate the calculations, a diagonal, not a full-covariance GMM UBM was used. The total amount of speakers in the training databases was about 4000.

The i-vector extractor was trained on more than 60000 telephone and microphone recordings from the NIST 1998-2010 comprising more than 4000 speakers' voices. The main expression defining the factor analysis of the GMM parameters with the aim of lowering data dimensionality is given below:

$$\mu = m + T\omega + \varepsilon \quad (1)$$

where μ is the supervector of the GMM parameters of the speaker model, m is the supervector of the UBM parameters, T is the matrix defining the basis in the reduced feature space, ω is the i-vector in the reduced feature space, $\omega \in N(0,1)$, ε is the error vector.

LDA matrix was trained on the same data from the NIST 1998-2010. In our current work, two speech databases were used to form the pairs: one contains Russian speech of Tajiks, the other contains Tajik speech of Tajiks. All the comparisons were cross-channel because it simulates a real forensic evidence situation. In total more than 121,000 pairs of recordings were automatically compared:

TABLE I
The Number of Automatically Compared Target-Target and Target-Imposter Pairs of Recordings Made Up of Two Speech Databases

Number of comparisons	Russian speech of Tajiks	Tajik speech of Tajiks
Target-target	401	394
Target-imposter	60481	60488

The obtained results were sorted by pseudo-probability value P calculates with the following formula:

$$P = (FR - FA)/2 + 50 \quad (2)$$

where FR is the false positive error, FA is the false negative error.

40 pairs were selected and given to the expert for comparison: 20 pairs contained Russian speech of Tajiks and 20 contained Tajik speech of Tajiks. In each 20 pairs, there were 10 target-target and 10 target-imposter pairs.

B. Expert

The expert who participated in the experiment is a specialist in applied phonetics with five years' experience of making identification comparisons. The expert did not know the correct answers or the ratio of target-target and target-imposter pairs in each 20 comparisons.

C. Experiment

The experiment was carried out according to the acoustic-spectrographic method described above. It should be noted that searching for such comparable utterances implies that the expert knows the language spoken in the recording. In our case, since the expert did not know the Tajik language, the same text read in Tajik was used to detect the comparable vowel realizations correctly. In Russian, the same text was read as well.

The expert analyzed formants of 5 types of vowels that can be compared in Russian and Tajik: [a], [i], [u] (back), [e], [o]. Russian [ɨ] (mixed) and Tajik [ɵ:] (central) were excluded for the reason that they could use bias the statistics because of their uniqueness.

The expert analyzed F1, F2, and F3 due to the fact that F4 was often beyond the spectral band, and the number of cases when F4 was visible was not representative.

The experiment was carried out using the audio forensic sound editor and visualizer SIS II developed in Speech Technology Center.

IV. Results

The total error rates are predictably high in both cases. This can be explained by the method of selecting the material: the most difficult pairs were selected, the way it is done in NIST competitions, therefore the error rates correspond to NIST HASR 2010 error rates [10].

TABLE II
ERROR RATES IN RUSSIAN SPEECH OF TAJIKS AND TAJIK SPEECH OF TAJIKS

	Russian speech of Tajiks	Tajik speech of Tajiks
Error rate	55 %	30 %

As can be seen from Table II, the error rate in non-native speech is higher than in Tajik speech of Tajiks. This difference between error rates corresponds to a greater degree of the difference between FA values. Fig. 1 demonstrates the rates FR and FA in both cases.

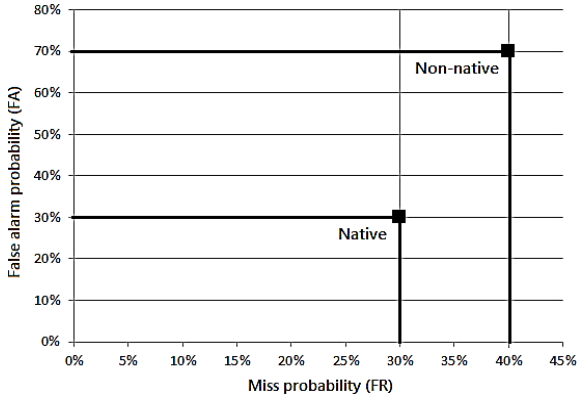

Fig. 1 The rates of FR and FA of foreign speech and native speech comparisons

A. Type I Errors

FR values can be considered similar in terms of the total quantity of ten target-target pairs. The average difference between formant frequency values is the same for both types of material (native speech and non-native speech): 7 %. This means that there are no dramatic deviations the expert should take into account when he or she makes a decision that two voices belong to different speakers in case of speech in a foreign language. Therefore, an expert can make false rejection errors with the same probability in cases of both types of material.

B. Type II Errors

It was more interesting to analyze the type II error because the difference between FA values in the case of native and non-native speech is significant (30 % for Tajik speech of Tajiks and 70 % for Russian speech of Tajiks). This is due to higher between-speaker variation of formant frequency values in the case of non-native speech. Between-speaker variation of formant frequency values was obtained by calculating formant frequency values in target-imposter pairs. The histograms presented in Fig. 2 show the average difference between the values of F1, F2, and F3 for each analyzed vowel type.

As can be seen from Fig. 2, in some cases the average difference between formant frequency values is lower in Tajik speech of Tajiks: F1 for [i] and [u], F2 for [a] and F3 for [i]. Still, in most cases it is lower in Russian speech of Tajiks. We can also see that on average (Fig. 3) the difference between all the formant frequencies is lower in non-native speech for almost all vowel types (except [i]).

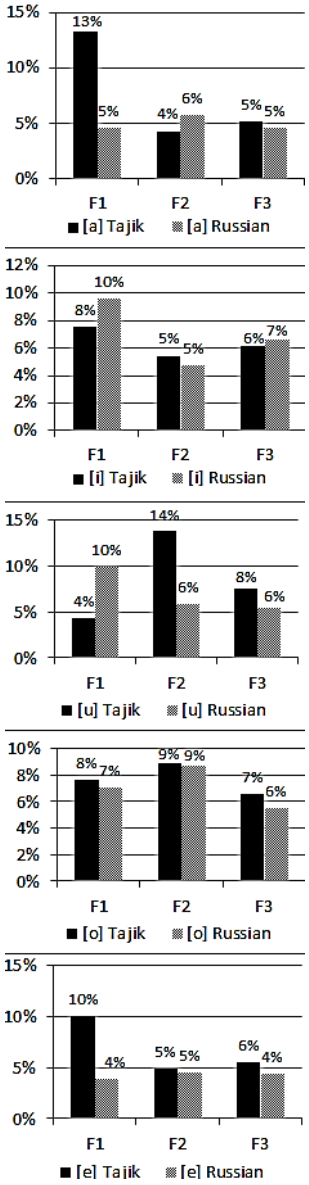

Fig. 2 Average differences between formant frequency values for each analyzed vowel type in Russian speech of Tajiks and Tajik speech of Tajiks

Therefore, the higher FA rate can be explained by the above-mentioned phenomenon of lower between-speaker variation in non-native speech.

V. CONCLUSION

We can conclude that an expert should pay particular attention to vowel formant frequency values when comparing recordings containing speech in a foreign language by the acoustic-spectrographic method. Therefore, the proposed hypothesis was confirmed. Lower between-speaker variability leads to higher error rate for this kind of material. Besides, because of the higher error rate the expert should have lower confidence in the method in case of non-native speech and probably rely more on other identification methods he or she uses in every single comparison.

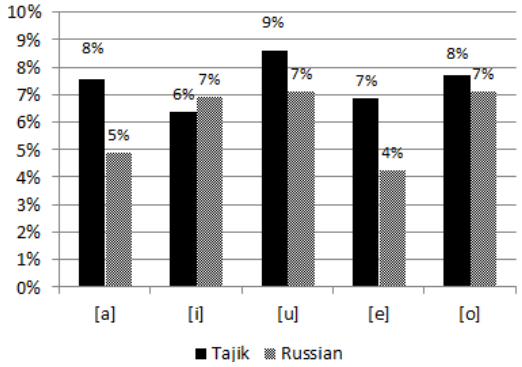

Fig. 3 Average differences between all formant frequency values for each analyzed vowel type in Russian speech of Tajiks and Tajik speech of Tajiks

This research was a pilot study, and it was carried out for a particular case of foreign speech, Russian speech of Tajiks. However, the trend we discovered can appear in other cases of the same type. The problem of forensic comparisons of speech in a non-native language is highly important nowadays and needs deeper exploration. We plan to proceed with research in this field that will involve more experts and a larger selection of difficult pairs of recordings that may contain spontaneous speech. This research is aimed at developing recommendations for experts who make forensic comparisons of recordings containing speech in a non-native language.

This work was partially financially supported by the Government of the Russian Federation, Grant 074-U01.

REFERENCES

[1] T. I. Goloshchapova, Yu. A., Elemeshina, "Expert-Linguistic Methods for Identification of Tajik Native Speakers," in *The 20th International Science Conference on Informatization and Information Protection of Law Enforcement*, 2011, pp. 402-405.

[2] S. L. Koval, P. I. Zubova, "Speaker identification by his voice and speech on the basis of complex analysis of phonograms," *Theory and practice of forensic expertise*, № 3 (7), pp. 68-76, Dec. 2007.

[3] U. Bhattacharjee, K. Sarmah, "Speaker Verification Using Acoustic and Prosodic Features." *Advanced Computing International Journal*, vol. 4, 1, pp. 45-51, Jan. 2013.

[4] A. R. Butcher, "Forensic Phonetics: Issues in Speaker Identification Evidence," in *Inaugural International Conference of the Institute of Forensic Studies*, 2002, pp. 3-5.

[5] K. Simonchik, T. Pekhovsky, A Shulipa, A. Afanasyev, "Supervised Mixture of PLDA Models for Cross-Channel Speaker Verification," in *Interspeech-2012*, http://speechpro-usa.com/files/filefield_stats/1798/1171/0/bf1ea2362f49c270c85a812b7b8e8311

[6] The NIST year 2010 speaker recognition evaluation plan, http://www.itl.nist.gov/iad/mig/tests/spk/2010/NIST_SRE10_evalplan.r6.pdf

[7] S. Aleinik, Yu. Matveev, A. Raev, "Method of Evaluation of Speech signal clipping level," *Scientific and Technical Journal of Information Technologies, Mechanics, and Optics*, vol. 79, 3, pp. 79-83, May 2012.

[8] Y. Matveev, K. Simonchik, "The Speaker Identification System for the NIST SRE 2010," in *The 20th International Conference on Computer Graphics and Vision, GraphiCon*, 2010, pp. 315-319.

[9] A. Kozlov, O. Kudashev, Yu. Matveev, T. Pekhovsky, K. Simonchik, A. Shulipa, "SVID Speaker Recognition System for NIST SRE 2012," in *Speech and Computer: Lecture Notes in Computer Science*, 2013, vol. 8113, pp. 278-285.

[10] C. S. Greenberg, A. F. Martin, L. Brandschain, J. P. Campbell, Ch. Cieri, G. R. Doddington, J. J. Godfrey, "Human Assisted Speaker Recognition in NIST SRE10," in *Odyssey*, 2010, pp. 180-185.