Usage of HMM-Based Speech Recognition Methods for Automated Determination of a Similarity Level Between Languages

Ansis Ataols Bērziņš(✉)

University of Latvia, Maskavas 54, Riga LV-1003, Latvia
ansis@latnet.lv

Abstract. The problem of automated determination of language similarity (or even defining of a distance on the space of languages) could be solved in different ways – working with phonetic transcriptions, with speech recordings or both of them. For the recordings, we propose and test a HMM-based one: in the first part of our article we successfully try language detection, afterwards we are trying to calculate distances between HMM-based models, using different metrics and divergences. The Kullback-Leibler divergence is the only one we got good results with – it means that the calculated distances between languages correspond to analytical understanding of similarity between them. Even if it does not work very well, the conclusion is that this method is usable, but usage of some other methods could be more rational.

Keywords: Distance between languages · Hidden Markov models · Kullback-Leibler divergence

1 Introduction

For some time already we have been searching for various methods for assessing proximity of natural idioms. An idiom is a common name for language varieties, regardless of their exact status [1]. In this article we use the terms "idiom" and "language" in a broad sense of the words, that is, the meaning includes tongue, dialect, language etc. Since the initial and main realization of an idiom is its oral form, we accept its existence as a prerequisite. The presence of a written form is not essential.

The problem of determining proximity or remoteness of idioms is of great practical importance for determining a degree of independence or non-independence of a language, in distinguishing languages and dialects, in clarifying a place of an idiom in language families and groups, in improving information modeling of cognitive processes. Scientifically, identifying proximity of idioms is a problem of linguistic taxonomy, which is trying to develop objective, purely linguistic tools for determining whether two close idioms are dialects or different languages – and this question already goes beyond linguistics into fields of social and political sciences. For example, in the context of the linguistic realities of Latvia, it is important to find out whether Latgalian is an independent language or a dialect of the Latvian language.

D. Ustalov et al. (Eds.): AINL 2019, CCIS 1119, pp. 113–125, 2019.
https://doi.org/10.1007/978-3-030-34518-1_8

Already a lot of research has been done on measuring distances between languages and dialects – mostly orthographical text data is used [9], in some cases – phonetic transcription of speech [8, 10, 11], even rarer – speech recordings, for example – by prosody [7]. In many cases fixed lexicons are used. The novelty of our research is the usage of full recordings of spontaneous speech for statistical models' building: turns out that these models are characterizing languages good enough to obtain distances between them.

For a long time hidden Markov models have been widely used for speech recognition tasks. This method is language-dependent because it is based on a dictionary or lexicon. The basic idea is that for every language's word a statistical model is created, based on a sufficient number of recordings of this word (must include variations of speakers, speed, intonation, context etc.). When a system needs to recognize a speech sample, it is first divided into fragments – each fragment is a single word. The task of splitting is not trivial, because in spontaneous speech there are often no clear breaks between words. In such cases the so-called phonotactics, i.e. knowledge about possible sounds' combinations in a given language, are most often used. In languages with many morphological forms, one can also try to separate a lexical part of a word (root) and the morphological part (in most cases – the end): in this case dictionary's size is smaller (contains only basic forms), but the programming of the software is more complex.

After that by Viterbi algorithm the closest, "most similar", most probable, hidden Markov model of the vocabulary is found for each fragment, and the name to which it corresponds is recognized as recognition of a given fragment of speech.

For more details on the method, as well as explanation and characterization of Hidden Markov models built on speech recordings, see, for example, [2] and [3].

Anyway, it is clear that a HMM-based speech recognition system will divide a speech into units, and the language and purpose depends only of their subtlety – whether it be words, syllables, phonemes or word groups. Thus, in terms of speech recognition, these above-mentioned units will be objects that will be described by hidden Markov models (or "words").

Unlike speech recognition we are interested not to detect and to transcribe speech units, but to evaluate languages as such and to determine a distance between them. Therefore, it would be logical to choose longer units of speech as HMM objects, which will characterize the language as a whole. In this case, the creation or "training" of a HMM should not take place on a given word (or its set or component), but on recordings of the whole language. Since, of course, no one can pronounce all the words and their combinations, one should at least strive for such a comprehension. We decided that it could be done by selecting an informant (=her/his recordings) as the object of HMM (if there are several recordings, they should be combined into one). Thus, the hidden Markov model of a language could be created on a sufficiently large (so as to ensure that it's speaker-independent) selection of informants or speakers. In order to be as close as possible to a live, natural language, recordings must be freely chosen, that is, they may be expeditions' recordings of spontaneous speech.

2 Data

Undoubtedly, such a method is applicable to any spoken languages (as we have repeatedly pointed out – we mean languages in a broad sense, including those without written form). However, as Latvian dialects were more accessible to us, we decided first to be based on them.

In year 2008 we have been collected our own spontaneous speech recordings of five Latvian dialects (recorded by the author of this article in folk-lore/linguistic expeditions) in Latgalia and Courland, four of them – Latgalian (Vileks, Baļtinova, Rudzātys and Auleja), and one – Couronian (Dundag) (Fig. 1).

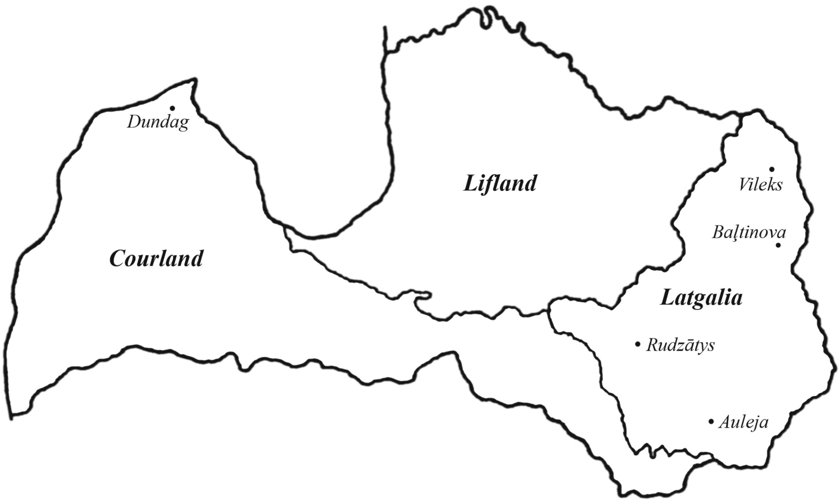

Fig. 1. The recorded dialects on the map of Latvia.

All recordings were collected in accordance with high principles of gathering [4], it means, all records were uniformed, recorded with the same type of hardware (a dynamic one-way microphone fixed on heads of speakers was used), an external noise was minimized as far as possible. All entries were manually cleared – i.e., all other voices and sounds were cut out, leaving only the speech of the main speaker. Recording technical quality was 44.1 kHz/16 bit.

All informants (Table 1) were asked to tell their life stories: about parents, grandparents, brothers, sisters, children, other family members, school, work, weddings, farm, military service, etc. It means the lexicon used by the informants was traditional and homogeneous.

Table 1. Characteristics of recordings used in the experiment.

Dialect	Minutes collected	Number of informants	Including male	Including female
Auleja	95	14	8	6
Baļtinova	140	23	9	14
Dundaga	161	17	4	13
Rudzātys	246	28	11	17
Vileks	238	30	11	19

3 Experiment

In fact, several experiments were carried out to find out and test the proposed method. They were all implemented by the help of HTK package [5], i.e. there was no need to program the algorithms and even to study their implementation in the package, since it is recognized among speech researchers worldwide. Of course, some scripts were developed for data processing and automation purposes.

Initially we would like to formalize the algorithm of our experiments step-by-step:

(1) speech samples of languages (dialects) we wanted to compare were selected;
(2) for each language a Hidden Markov model was created, using selected samples (full recordings were used, without any cutting);
(3) different measures (metrics, divergences) were tried to measure distance between newly created models pair by pair;
(4) the numerical results of each distance were compared with analytical and intuitive understanding of how close or far the analysed languages are;
(5) for each distance conclusions about applicability of such a distance were drawn out.

The idea of the first two experiments was language identification task by HMM created on long recordings of different speakers of several languages.

The first experiment was carried out with a read speech: the same text read by the same person in three languages – Latvian, Latgalian and Russian. Four recordings were recorded in each language: three were read at medium speed and one – at accelerated; length of each of the recordings – 1 to 2 min. For each language on all the three medium speed's speech recordings hidden Markov model was created. After that with the HVite utility (the implementation of the Viterbi algorithm in the HTK package) the nearest model for each of the high-speed speech recordings was founded. With a small number of Gaussian mix components (so-called "mods") the results were unsatisfactory, but with four and above worked properly – the high-speed speech recordings' languages were detected flawlessly (Table 2).

The positive results of this experiment motivated us to do the next one, this time on the real data of our research.

Table 2. The results of the first experiment.

Language	Mixtures				
	0	2	4	8	16
Of the recording	*Detected*				
Latgalian	Latgalian	Latgalian	Latgalian	Latgalian	Latgalian
Latvian	Latgalian	Latvian	Latvian	Latvian	Latvian
Russian	Latgalian	Latgalian	Russian	Russian	Russian

We chose two from our recorded dialects – Rudzātys and Vileks, both Latgalian, but from opposite sides of Latgalia: Northeast and Southwest. Thus, the chosen languages were very close (and it, of course, reinforces the importance of results in a case of a positive outcome), but at the same time far enough to be sure that differences will not be smothered by social contacts of speakers. From each language we randomly chose eight female[1] informants, those eight were randomly divided into two subgroups: five for model creation and three for testing. The results were identical to the results of the previous experiment: in the case of a small number of mods, languages were detected erroneously (in different ways, without understandable consequences), but in case of four or more – flawlessly (Table 3).

Table 3. The results of the second experiment.

Language	Mixtures				
	0	2	4	8	16
Of the recording	*Detected*				
Vileks	Rudzātys	Rudzātys	Vileks	Vileks	Vileks
Vileks	Vileks	Vileks	Vileks	Vileks	Vileks
Vileks	Vileks	Vileks	Vileks	Vileks	Vileks
Rudzātys	Rudzātys	Vileks	Rudzātys	Rudzātys	Rudzātys
Rudzātys	Rudzātys	Vileks	Rudzātys	Rudzātys	Rudzātys
Rudzātys	Rudzātys	Rudzātys	Rudzātys	Rudzātys	Rudzātys

Thus, we can conclude that our hypothesis of the possibility of training HMMs on full-size recordings to describe language as such was confirmed. We assumed that once it works in recognition tasks, i.e., the language of other recordings is correctly determined by such models, it should also work in determination of the distance between languages,

[1] We chose women because we collected more female voice speech data in our expeditions – apparently because women live longer [12] and are more talkative (at least by our observations, although in research their predominance of daily word use does not meet thresholds for statistical significance, e.g., [13, 14]).

Table 4. Euclidean metrics for the mean value vectors of the read speech.

Distance: **eikl**

	lg1	lg2	lg3	lg4	lv1	lv2	lv3	lv4	ru1	ru2	ru3	ru4
lg1	0	22.749	22.278	27.159	22.724	22.269	22.593	25.059	22.053	24.055	21.753	25.956
lg2	22.749	0	21.883	28.395	24.354	22.651	21.458	18.153	15.944	23.391	24.896	20.777
lg3	22.278	21.883	0	24.88	23.548	21.228	20.315	20.394	21.271	24.631	24.973	21.522
lg4	27.159	28.395	24.88	0	26.289	26.914	29.626	26.387	27.78	28.796	27.597	25.745
lv1	22.724	24.354	23.548	26.289	0	18.682	19.487	22.612	19.864	20.375	24.491	23.907
lv2	22.269	22.651	21.228	26.914	18.682	0	15.721	21.131	21.135	22.921	23.931	24.765
lv3	22.593	21.458	20.315	29.626	19.487	15.721	0	21.265	21.052	21.434	22.388	23.698
lv4	25.059	18.153	20.394	26.387	22.612	21.131	21.265	0	17.247	22.727	24.383	21.04
ru1	22.053	15.944	21.271	27.78	19.864	21.135	21.052	17.247	0	21.729	22.011	21.907
ru2	24.055	23.391	24.631	28.796	20.375	22.921	21.434	22.727	21.729	0	17.821	24.128
ru3	21.753	24.896	24.973	27.597	24.491	23.931	22.388	24.383	22.011	17.821	0	24.447
ru4	25.956	20.777	21.522	25.745	23.907	24.765	23.698	21.04	21.907	24.128	24.447	0

i.e. we could define a distance between languages as a distance between HMMs of these languages.

That's why we decided to create HMMs for all the five of our dialects and define different types of metrics in their space. After that we started the second part of our experiments – to try out different distance measures on newly created hidden Markov models pair by pair.

4 Euclidean Metrics and Its Improvements

Initially, we decided to try our luck with the well-known metric – Euclidean. Then, the choice was made as for the data (characterizing the distribution) that would be dimensions of our metric space. It seemed reasonable to use mean value vectors (model includes mean, variance and weight vectors).

Firstly we made distance calculations for the above mentioned Latvian/ Latgalian/Russian read speech. We calculated Euclidean metrics, normalized Euclidean metrics (normalized by the first, second, and both arguments) and Gordian metrics.

In the Tables 4, 5 and 6 are used notation: *lg* – Latgalian, *lv* – Latvian, *ru* – Russian; the following number is a serial number of the recording of this particular language, for example, *ru2* means the second recording of Russian speech.

In case of correct distances one should expect that distances between speech samples of the same language are smaller, between Latvian and Latgalian – medium, between Russian and Latgalian – bigger, and between Russian and Latvian – biggest ones. However, for all the three metrics, it can be seen that the distances are very similar, and at the same time they are "jumping" – having unpredictable changes, that makes possible, that intuitively closer languages have larger distances and vice versa.

We carried out this experiment on our dialects' speech samples too.

Unfortunately, the program HERest from the HTK package, which performs a recalculation of HMM parameters using the Baum-Welch algorithm[2], obviously has a fault – at a larger number of input files, it displays an error message that approximation cannot

[2] *This program is used to perform a single re-estimation of the parameters of a set of HMMs using an embedded training version of the Baum-Welch algorithm. Training data consists of one or more utterances each of which has a transcription in the form of a standard label file (segment boundaries are ignored). For each training utterance, a composite model is effectively synthesized by concatenating the phoneme models given by the transcription.* [5]

Table 5. Gordian metrics for the mean value vectors of the read speech.

Distance: **zord**

	lg1	lg2	lg3	lg4	lv1	lv2	lv3	lv4	ru1	ru2	ru3	ru4
lg1	0	25.649	24.059	31.27	29.15	34.27	30.701	25.499	22.682	27.701	28.456	25.113
lg2	25.649	0	23.029	30.55	25.91	28.666	27.372	24.873	23.724	25.798	25.049	24.265
lg3	24.059	23.029	0	30.64	26.5	26.061	23.47	24.493	23.015	25.662	27.592	24.542
lg4	31.27	30.55	30.64	0	26.098	23.247	32.635	28.445	27.157	28.48	29.721	27.695
lv1	29.15	25.91	26.5	26.098	0	29.649	27.779	26.027	23.549	26.36	25.267	25.333
lv2	34.27	28.666	26.061	23.247	29.649	0	20.107	22.198	22.967	31.48	26.833	23.877
lv3	30.701	27.372	23.47	32.635	27.779	20.107	0	26.123	25.054	27.912	31.831	24.4
lv4	25.499	24.873	24.493	28.445	26.027	22.198	26.123	0	23.574	24.751	25.971	21.272
ru1	22.682	23.724	23.015	27.157	23.549	22.967	25.054	23.574	0	23.333	24.553	23.309
ru2	27.701	25.798	25.662	28.48	26.36	31.48	27.912	24.751	23.333	0	27.435	27.425
ru3	28.456	25.049	27.592	29.721	25.267	26.833	31.831	25.971	24.553	27.435	0	33.328
ru4	25.113	24.265	24.542	27.695	25.333	23.877	24.4	21.272	23.309	27.425	33.328	0

Table 6. Normalized by both arguments Euclidean metrics for the mean value vectors of the read speech.

Distance: **norm**

	lg1	lg2	lg3	lg4	lv1	lv2	lv3	lv4	ru1	ru2	ru3	ru4
lg1	0	0.32	0.315	0.377	0.322	0.306	0.315	0.355	0.314	0.339	0.307	0.359
lg2	0.32	0	0.305	0.395	0.347	0.314	0.299	0.254	0.221	0.329	0.35	0.29
lg3	0.315	0.305	0	0.344	0.336	0.301	0.286	0.289	0.3	0.354	0.357	0.3
lg4	0.377	0.395	0.344	0	0.372	0.376	0.414	0.366	0.389	0.402	0.389	0.349
lv1	0.322	0.347	0.336	0.372	0	0.265	0.277	0.323	0.286	0.289	0.355	0.333
lv2	0.306	0.314	0.301	0.376	0.265	0	0.221	0.302	0.301	0.322	0.341	0.347
lv3	0.315	0.299	0.286	0.414	0.277	0.221	0	0.299	0.298	0.304	0.316	0.332
lv4	0.355	0.254	0.289	0.366	0.323	0.302	0.299	0	0.246	0.325	0.348	0.291
ru1	0.314	0.221	0.3	0.389	0.286	0.301	0.298	0.246	0	0.308	0.315	0.303
ru2	0.339	0.329	0.354	0.402	0.289	0.322	0.304	0.325	0.308	0	0.255	0.337
ru3	0.307	0.35	0.357	0.389	0.355	0.341	0.316	0.348	0.315	0.255	0	0.343
ru4	0.359	0.29	0.3	0.349	0.333	0.347	0.332	0.291	0.303	0.337	0.343	0

Table 7. Euclidean metrics for the mean value vectors of the spontaneous dialect speech.

Distance: **eikl**

	auleja	auleja_m	baltinova	baltinova_m	dundag	dundag_m	rudzati	rudzati_m	vileks	vileks_m
auleja	0	15.164	15.411	14.949	16.812	15.383	16.907	15.66	17.383	15.738
auleja_m	15.164	0	14.013	12.847	11.325	9.937	12.028	11.507	14.649	13.961
baltinova	15.411	14.013	0	12.979	12.09	12.007	13.375	12.402	13.188	11.972
baltinova_m	14.949	12.847	12.979	0	12.199	12.534	11.23	12.793	14.621	12.862
dundag	16.812	11.325	12.09	12.199	0	10.568	9.744	11.013	12.076	11.919
dundag_m	15.383	9.937	12.007	12.534	10.568	0	12.897	12.294	13.292	13.149
rudzati	16.907	12.028	13.375	11.23	9.744	12.897	0	10.752	12.595	11.123
rudzati_m	15.66	11.507	12.402	12.793	11.013	12.294	10.752	0	13.048	11.574
vileks	17.383	14.649	13.188	14.621	12.076	13.292	12.595	13.048	0	12.206
vileks_m	15.738	13.961	11.972	12.862	11.919	13.149	11.123	11.574	12.206	0

be calculated: *WARNING [−7324] StepBack: File [path] - bad data or over pruning*. Such a problem should occur if the recording is technically poor or has some other fault. However, it is interesting that for a same file this error could appear with a larger number of files, but not appear with a smaller one – hence it does not depend on the file quality, but on something else. This leads to the conclusion that this is a fault of the program, and the only way to avoid it is to bypass it. As we simply did not want to skip some of the files, we decided to divide the voices of men and women into separate groups – there were fewer files in each group and HERest stopped crashing. Thus, the experiment

Table 8. Gordian metrics for the mean value vectors of the spontaneous dialect speech.

Distance: **zord**

	auleja	auleja_m	baltinova	baltinova_m	dundag	dundag_m	rudzati	rudzati_m	vileks	vileks_m
auleja	0	23.823	24.657	19.026	24.251	24.331	27.276	28.447	28.626	27.885
auleja_m	23.823	0	18.64	15.576	22.261	21.345	15.4	19.894	22.295	18.611
baltinova	24.657	18.64	0	13.863	10.268	12.964	12.453	10.988	18.701	12.775
baltinova_m	19.026	15.576	13.863	0	14.831	12.15	14.599	20.069	18.922	14.786
dundag	24.251	22.261	10.268	14.831	0	9.709	11.357	9.778	13.147	10.799
dundag_m	24.331	21.345	12.964	12.15	9.709	0	12.145	12.324	14.284	12.741
rudzati	27.276	15.4	12.453	14.599	11.357	12.145	0	9.958	13.431	13.08
rudzati_m	28.447	19.894	10.988	20.069	9.778	12.324	9.958	0	14.896	15.981
vileks	28.626	22.295	18.701	18.922	13.147	14.284	13.431	14.896	0	15.691
vileks_m	27.885	18.611	12.775	14.786	10.799	12.741	13.08	15.981	15.691	0

Table 9. Normalized by both arguments Euclidean metrics for the mean value vectors of the spontaneous dialect speech.

Distance: **norm**

	auleja	auleja_m	baltinova	baltinova_m	dundag	dundag_m	rudzati	rudzati_m	vileks	vileks_m
auleja	0	0.258	0.26	0.253	0.291	0.266	0.287	0.263	0.299	0.264
auleja_m	0.258	0	0.258	0.233	0.208	0.183	0.219	0.207	0.267	0.251
baltinova	0.26	0.258	0	0.238	0.229	0.227	0.249	0.229	0.246	0.219
baltinova_m	0.253	0.233	0.238	0	0.225	0.233	0.203	0.231	0.267	0.231
dundag	0.291	0.208	0.229	0.225	0	0.203	0.182	0.205	0.228	0.22
dundag_m	0.266	0.183	0.227	0.233	0.203	0	0.243	0.23	0.252	0.244
rudzati	0.287	0.219	0.249	0.203	0.182	0.243	0	0.196	0.232	0.201
rudzati_m	0.263	0.207	0.229	0.231	0.205	0.23	0.196	0	0.238	0.209
vileks	0.299	0.267	0.246	0.267	0.228	0.252	0.232	0.238	0	0.222
vileks_m	0.264	0.251	0.219	0.231	0.22	0.244	0.201	0.209	0.222	0

Table 10. Euclidean metrics for the mean value vectors divided by the variances, for the read speech.

Distance: **eikl**

	lg1	lg2	lg3	lg4	lv1	lv2	lv3	lv4	ru1	ru2	ru3	ru4
lg1	0	9.441	8.671	9.969	8.952	8.705	8.979	9.591	9.852	9.114	9.58	8.856
lg2	9.441	0	7.855	9.83	8.798	8.627	8.993	7.636	5.875	10.662	10.212	8.691
lg3	8.671	7.855	0	9.169	9.143	8.936	8.638	8.615	7.171	11.326	10.23	9.11
lg4	9.969	9.83	9.169	0	10.164	10.863	10.687	9.185	9.995	11.013	10.384	7.901
lv1	8.952	8.798	9.143	10.164	0	6.813	7.696	9.991	8.493	7.393	9.926	9.736
lv2	8.705	8.627	8.936	10.863	6.813	0	7.505	9.907	8.572	8.074	10.406	9.698
lv3	8.979	8.993	8.638	10.687	7.696	7.505	0	9.658	8.525	10.063	9.034	9.947
lv4	9.591	7.636	8.615	9.185	9.991	9.907	9.658	0	8.008	11.338	11.31	9.008
ru1	9.852	5.875	7.171	9.995	8.493	8.572	8.525	8.008	0	10.555	10.17	9.875
ru2	9.114	10.662	11.326	11.013	7.393	8.074	10.063	11.338	10.555	0	9.065	10.227
ru3	9.58	10.212	10.23	10.384	9.926	10.406	9.034	11.31	10.17	9.065	0	9.77
ru4	8.856	8.691	9.11	7.901	9.736	9.698	9.947	9.008	9.875	10.227	9.77	0

became larger and probably more interesting, but it also has one drawback – we will not be able to compare directly its results with results of other methods.

Notation used in the Tables 7, 8 and 9: the name of the dialect without any additions means model built on the recordings of women voices, with "_m" at the end means model built on men voices.

As we can see, all the distances here are "dancing" – "men" of the same language sometimes are farther than "women" of other language, intuitively close languages sometimes appear farther than distant ones.

At the suggestion of Professor, Dr. habil. math. Aivars Lorencs, we decided to try the same metrics, but for the mean values divided by the variances, that is, the more volatile

Table 11. Gordian metrics for the mean value vectors divided by the variances, for the read speech.

Distance: **zord**

	lg1	lg2	lg3	lg4	lv1	lv2	lv3	lv4	ru1	ru2	ru3	ru4
lg1	0	38.01	36.904	34.635	34.366	32.003	38.333	35.821	37.778	36.224	36.133	37.656
lg2	38.01	0	19.786	28.913	37.509	35.146	29.962	30.685	14.212	39.367	37.9	20.78
lg3	36.904	19.786	0	24.124	34.678	32.315	22.317	27.772	22.794	36.536	35.864	20.814
lg4	34.635	28.913	24.124	0	37.57	35.207	30.948	32.804	31.921	39.428	38.578	15.005
lv1	34.366	37.509	34.678	37.57	0	12.254	29.85	39.77	38.421	14.37	37.356	35.936
lv2	32.003	35.146	32.315	35.207	12.254	0	39.146	37.407	36.058	23.515	36.819	33.574
lv3	38.333	29.962	22.317	30.948	29.85	39.146	0	30.615	21.065	39.393	32.469	36.219
lv4	35.821	30.685	27.772	32.804	39.77	37.407	30.615	0	31.378	41.628	37.693	34.669
ru1	37.778	14.212	22.794	31.921	38.421	36.058	21.065	31.378	0	40.28	38.443	21.452
ru2	36.224	39.367	36.536	39.428	14.37	23.515	39.393	41.628	40.28	0	39.214	37.795
ru3	36.133	37.9	35.864	38.578	37.356	36.819	32.469	37.693	38.443	39.214	0	31.556
ru4	37.656	20.78	20.814	15.005	35.936	33.574	36.219	34.669	21.452	37.795	31.556	0

Table 12. Normalized by both arguments Euclidean metrics for the mean value vectors divided by the variances, for the read speech.

Distance: **norm**

	lg1	lg2	lg3	lg4	lv1	lv2	lv3	lv4	ru1	ru2	ru3	ru4
lg1	0	0.811	0.706	0.914	0.779	0.744	0.784	0.886	0.779	0.82	0.809	0.871
lg2	0.811	0	0.795	0.918	0.816	0.794	0.835	0.647	0.581	0.848	0.86	0.817
lg3	0.706	0.795	0	0.882	0.812	0.733	0.76	0.814	0.718	0.879	0.839	0.883
lg4	0.914	0.918	0.882	0	0.914	0.944	0.984	0.884	0.904	0.896	0.894	0.819
lv1	0.779	0.816	0.812	0.914	0	0.705	0.765	0.88	0.786	0.786	0.893	0.912
lv2	0.744	0.794	0.733	0.944	0.705	0	0.653	0.839	0.759	0.792	0.865	0.912
lv3	0.784	0.835	0.76	0.984	0.765	0.653	0	0.851	0.799	0.85	0.866	0.947
lv4	0.886	0.647	0.814	0.884	0.88	0.839	0.851	0	0.701	0.894	0.952	0.861
ru1	0.779	0.581	0.718	0.904	0.786	0.759	0.799	0.701	0	0.84	0.833	0.884
ru2	0.82	0.848	0.879	0.896	0.786	0.792	0.85	0.894	0.84	0	0.707	0.877
ru3	0.809	0.86	0.839	0.894	0.893	0.865	0.866	0.952	0.833	0.707	0	0.871
ru4	0.871	0.817	0.883	0.819	0.912	0.912	0.947	0.861	0.884	0.877	0.871	0

Table 13. Euclidean metrics for the mean value vectors divided by the variances, for the spontaneous dialect speech.

Distance: **eikl**

	auleja	auleja_m	baltinova	baltinova_m	dundag	dundag_m	rudzati	rudzati_m	vileks	vileks_m
auleja	0	14.726	13.489	11.752	17.629	16.899	14.602	22.076	18.021	13.053
auleja_m	14.726	0	8.924	11.803	10.194	9.828	9.37	18.749	12.753	15.165
baltinova	13.489	8.924	0	10.437	11.001	10.238	9.044	18.947	11.307	13.747
baltinova_m	11.752	11.803	10.437	0	13.819	13.146	9.065	20.29	14.431	11.516
dundag	17.629	10.194	11.001	13.819	0	12.206	11.238	16.834	13.005	14.846
dundag_m	16.899	9.828	10.238	13.146	12.206	0	10.258	18.876	13.495	17.248
rudzati	14.602	9.37	9.044	9.065	11.238	10.258	0	18.01	12.474	13.45
rudzati_m	22.076	18.749	18.947	20.29	16.834	18.876	18.01	0	20.755	17.61
vileks	18.021	12.753	11.307	14.431	13.005	13.495	12.474	20.755	0	16.414
vileks_m	13.053	15.165	13.747	11.516	14.846	17.248	13.45	17.61	16.414	0

are values, the smaller is a weight – they are affecting less a value of the distance. The same notation as for Tables 4, 5, 6 and 7, 8, 9 are used (Tables 10, 11, 12, 13, 14 and 15).

As we can see, in any case, namely, for any data set and any metric, this improvement has not made results consistent.

That's why our conclusion is negative: we cannot define the distance in this way and should look for other ways to do it.

Table 14. Gordian metrics for the mean value vectors divided by the variances, for the spontaneous dialect speech.

Distance: **zord**

	auleja	auleja_m	baltinova	baltinova_m	dundag	dundag_m	rudzati	rudzati_m	vileks	vileks_m
auleja	0	103.891	104.791	57.378	109.976	125.074	95.018	101.771	111.753	57.455
auleja_m	103.891	0	13.618	46.512	23.73	21.183	13.159	30.749	54.44	46.436
baltinova	104.791	13.618	0	47.412	24.849	20.283	12.369	29.55	56.314	47.336
baltinova_m	57.378	46.512	47.412	0	52.597	67.695	37.639	44.393	59.371	26.323
dundag	109.976	23.73	24.849	52.597	0	22.929	22.274	22.269	45.355	52.521
dundag_m	125.074	21.183	20.283	67.695	22.929	0	30.056	30.528	54.702	67.619
rudzati	95.018	13.159	12.369	37.639	22.274	30.056	0	27.222	59.557	37.563
rudzati_m	101.771	30.749	29.55	44.393	22.269	30.528	27.222	0	40.731	44.316
vileks	111.753	54.44	56.314	59.371	45.355	54.702	59.557	40.731	0	54.298
vileks_m	57.455	46.436	47.336	26.323	52.521	67.619	37.563	44.316	54.298	0

Table 15. Normalized by both arguments Euclidean metrics for the mean value vectors divided by the variances, for the spontaneous dialect speech.

Distance: **norm**

	auleja	auleja_m	baltinova	baltinova_m	dundag	dundag_m	rudzati	rudzati_m	vileks	vileks_m
auleja	0	0.815	0.818	0.847	0.911	0.85	0.939	0.924	0.841	0.828
auleja_m	0.815	0	0.877	0.895	0.732	0.739	0.829	0.974	0.877	0.977
baltinova	0.818	0.877	0	0.879	0.865	0.898	0.974	1.018	0.872	0.898
baltinova_m	0.847	0.895	0.879	0	0.932	0.846	0.763	1.012	0.868	0.938
dundag	0.911	0.732	0.865	0.932	0	0.81	0.804	0.854	0.834	0.85
dundag_m	0.85	0.739	0.898	0.846	0.81	0	0.8	0.952	0.891	0.999
rudzati	0.939	0.829	0.974	0.763	0.804	0.8	0	0.936	0.872	0.907
rudzati_m	0.924	0.974	1.018	1.012	0.854	0.952	0.936	0	1.001	0.805
vileks	0.841	0.877	0.872	0.868	0.834	0.891	0.872	1.001	0	0.852
vileks_m	0.828	0.977	0.898	0.938	0.85	0.999	0.907	0.805	0.852	0

5 Kullback-Leibler Divergence

The most common assessment of HMM similarity is the Kullback-Leibler divergence, which the authors have been defined in their publication of 1951[3].

It is a mathematical expectation of a logarithmic difference between two probabilities distributions by the first distribution. So, naturally, it is not symmetrical, so it does not correspond to one of the axioms of metrics and is not a metric. Defining an arithmetic mean of divergence values of both directions often solves this problem.

Kullback-Leibler divergence was calculated with a slightly modified Python script written by Speech Lab of Technical University of Brno (Table 16).

At first glance, we can see a certain coherence in the results (e.g., the fact that Dundag looks further, or the fact that Baļtinova and Vileks is the closest pair), though, of course, the lack of symmetry and the separation of the voices of men and women is confusing and does not allow to analyze the results properly. Therefore, we decided to simplify them: first, to symmetrize the table by calculation of average arithmetic values and, secondly,

[3] *We are also concerned with the statistical problem of discrimination, by considering a measure of "distance" or "divergence" between statistical populations in terms of our measure of information. For the statistician two populations differ more or less according as to how difficult it is to discriminate between them with the best test. The particular measure we use has been considered by Jeffreys in another connection. He is primarily concerned with its use in providing an invariant density of a priory probability. A special case of this divergence is Mahalanobis' generalized distance.* [6]

Table 16. Kullback-Leibler divergence for the spontaneous dialect speech.

	Auleja, male	*Auleja, female*	*Baļtinova, male*	*Baļtinova, female*	*Dundag, male*	*Dundag, female*	*Rudzātys, male*	*Rudzātys, female*	*Vileks, male*	*Vileks, female*
Auleja,m.	0	2,77	2,99	3,13	4,31	3,69	5,12	2,91	3,75	3,80
Auleja, f.	2,83	0	5,17	5,39	12,75	6,98	7,42	4,42	4,59	5,44
Baļtinova, m.	4,17	4,46	0	3,08	7,14	4,28	5,54	2,95	2,78	3,31
Baļtinova, f.	3,90	4,64	2,99	0	6,21	3,25	7,97	2,34	4,29	2,22
Dundag, m.	2,32	3,92	3,17	3,18	0	2,96	6,51	3,22	4,26	3,37
Dundag, f.	2,91	3,43	2,87	2,45	3,55	0	5,69	2,05	3,86	2,87
Rudzātys, m.	3,29	2,79	2,91	4,01	4,13	3,67	0	2,71	2,63	4,84
Rudzātys ,f.	3,52	3,18	2,81	2,30	5,14	2,65	4,46	0	3,44	2,78
Vileks, m.	3,76	3,67	2,31	3,44	6,30	4,20	4,18	2,75	0	4,16
Vileks, f.	6,03	6,52	4,90	3,53	9,55	5,79	8,98	4,22	5,92	0

Table 17. Symmetrized Kullback-Leibler divergence for the spontaneous dialect speech (values rounded).

	Auleja	*Baļtinova*	*Dundag*	*Rudzātys*	*Vileks*
Auleja	0,00	4,23	5,04	4,08	4,70
Baļtinova	4,23	0,00	4,07	3,85	3,35
Dundag	5,04	4,07	0,00	4,13	5,03
Rudzātys	4,08	3,85	4,13	0,00	4,23
Vileks	4,70	3,35	5,03	4,23	0,00

to put together the male and female voices, also by taking the average arithmetic value (Table 17).

As we can see, this has brought all the values closer, which confirms that such a great range of values had other reasons than the qualities of languages. This, of course, is not good. However such similar values might reflect something – so let's look at them.

The distances of Auleja looks adequately: Dundag – the farthest, Rudzātys – the closest, Baļtinova closer than Vileks.

The results of Baļtinova could also be considered (Vileks very close, Rudzātys further) good if it were not for the unjustified Dundag's proximity to Auleja.

Even worse results for Rudzātys – Baļtinova appeared to be closer to Auleja, Dundag – closer to Vileks.

In contrast, Vileks looks very good – Baļtinova is the closest, then Rudzātys, then Auleja, and Dundag the farthest.

6 Discussion and Conclusions

Hidden Markov models, created on a set of long enough spontaneous speech recordings of a big enough number of different speakers of this language, are applicable for language detection tasks.

Euclidean metrics, Gordian metrics, and normalized by both arguments Euclidean metrics on the space of these models are not characterizing the relations between real objects the models are created for.

The (symmetrized) Kullback-Leibler divergence could be used as a distance between these HM models. It would be possible (and interesting) to try out the Jensen-Shannon distance and Jensen-Shannon divergence too, however, because the (symmetrized) Kullback-Leibler divergence works well enough, there is not a big need for that.

In general the method – HMM-based automated determination of a similarity level between languages – is usable. However, it is technically complex and the results are not fully reliable. Therefore, other methods, such as i-Vector, are more recommended for real use. By the word, we have been realized similar experiments based on more modern speech recognition technologies too, but these results are topic of other (future) publications.

References

1. Виноградов, В.А.: Идиом. Лингвистический энциклопедический словарь/Под ред. В.Н. Ярцевой, стр. 685. Советская энциклопедия, Москва (1990)
2. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 262–286 (1989)
3. Кушнир, Д.А.: Алгоритм формирования структуры эталона для пословного дикторо-независимого распознавания команд ограниченного словаря. Штучный інтелект № 3’2006, Київ (2006)
4. ბერზინი, ა.[berzini, a.] ინფორმაციის მოპოვების პრინციპები ფონოგრამების ავტომატური ანალიზისთვის[inp’ormats’iis mopovebis prints’ipebi p’onogramebis avtomaturi analizist’vis] = Принципы сбора информации для автоматизированного анализа фонограмм. ქართული ენა და თანამედროვე ტექნოლოგიები- 2011 [k’art’uli ena da t’anamedrove tek’nologiebi - 2011] стр. 39–46. მერიდიანი[meridiani], თბილისი[t’bilisi] (2011)
5. Young, S., et al.: The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department, Cambridge (2009)
6. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
7. Šimko, J., Suni, A., Hiovain, K., Vainio, M.: Comparing languages using hierarchical prosodic analysis. In: Proceedings of Interspeech 2017, pp. 1213–1217 (2017)
8. Nerbonne, J., Heeringa, W., van den Hout, E., van der Kooi, P., Otten, S., van de Vis, S.W.: Phonetic distance between Dutch dialects. In: CLIN VI, Papers from the Sixth CLIN Meeting. Antwerp: University of Antwerp, Center for Dutch Language and Speech, pp. 185–202
9. Tambovtsev, Y.: Phonological similarity between basque and other world languages based on the frequency of occurrence of certain typological consonantal features. Prague Bull. Math. Linguist. **79–80**, 121–126 (2003)

10. Berzinch, A.A.: La comparaison de typologie traditionnelle et de typologie phonolexique, basée sur la méthode des n-grammes, dans les dialectes baltes. Identification des langues et des variétés dialectales par les humains et par les machines. Paris: École National Supérieure des Télécommunications (2004)
11. Берзинь, А.У.: Измерение фономорфолексического расстояния между латышскими наречиями путём применения расстояния Вагнера-Фишера. Труды международной конференции. Диалог 2006. М.: Издательство РГГУ (2006)
12. Demogrāfija 2018: statistisko datu krājums. R.: Centrālā statistikas pārvalde (2018)
13. Mehl, M.R., Vazire, S., Ramírez-Esparza, N., Slatcher, R.B., Pennebaker, J.W.: Are women really more talkative than men? Science **317**(5832), 82 (2007). American Association for the Advancement of Science, Washington
14. Liberman M.: Sex-Linked Lexical Budgets. Language Log 2006/2007. http://itre.cis.upenn.edu/~myl/languagelog/archives/003420.html. Accessed 15 Sept 2019