Sequential Three-Way Decisions in Efficient Classification of Piecewise Stationary Speech Signals

Andrey V. Savchenko(✉)

Laboratory of Algorithms and Technologies for Network Analysis,
National Research University Higher School of Economics,
Nizhny Novgorod, Russian Federation
avsavchenko@hse.ru

Abstract. In this paper it is proposed to improve performance of the automatic speech recognition by using sequential three-way decisions. At first, the largest piecewise quasi-stationary segments are detected in the speech signal. Every segment is classified using the maximum a-posteriori (MAP) method implemented with the Kullback-Leibler minimum information discrimination principle. The three-way decisions are taken for each segment using the multiple comparisons and asymptotical properties of the Kullback-Leibler divergence. If the non-commitment option is chosen for any segment, it is divided into small subparts, and the decision-making is sequentially repeated by fusing the classification results for each subpart until accept or reject options are chosen or the size of each subpart becomes relatively low. Thus, each segment is associated with a hierarchy of variable-scale subparts (granules in rough set theory). In the experimental study the proposed procedure is used in speech recognition with Russian language. It was shown that our approach makes it possible to achieve high efficiency even in the presence of high level of noise in the observed utterance.

Keywords: Signal processing · Speech recognition · Three-way decisions · Sequential analysis · Granular computing · Kullback-Leibler divergence

1 Introduction

The mathematical model of the piecewise stationary stochastic (random) process [1,2] is widely used in many practical pattern recognition tasks including signal classification [3,4], computer vision [5] and speech processing [6]. One of the most popular approach to classify its realization (sample function) is based on the hidden Markov model (HMM), specially developed for recognition of the piecewise stationary signals [6]. In these methods an observed realization of stochastic process [7] is divided into stationary parts using a fixed scale time window (typically 20–30 ms) [1]. Next, the corresponding parts (segments) of

L. Polkowski et al. (Eds.): IJCRS 2017, Part II, LNAI 10314, pp. 264–277, 2017.
DOI: 10.1007/978-3-319-60840-2_19

the observation and all instances in the database are matched using such models of these segments, as the GMM (Gaussian Mixture Model), and the total similarity is estimated. The recent research has moved focus from GMMs to more complex classifiers based on the deep neural networks (DNN), which have established the state-of-the-art results for several multimedia recognition tasks [8,9]. The most impressive modern results are achieved with acoustic models based on long-short term memory (LSTM) recurrent neural networks trained with connectionist temporal classification [10]. Unfortunately, the run-time complexity of all these approaches is rather high, especially for large utterances, which contain many phones [6,11]. In practice the situation is even worse, because the segments are usually aligned using dynamic programming to deal with inaccurate segmentation.

It is known [1], that the speech signals are multi-scale in nature (vowel phones last for 40–400 ms while stops last for 3–250 ms). Hence, to improve classification performance, this paper explores the potential of sequential three-way decisions (TWD) [12], which has been recently used to speed-up the face recognition algorithms [13,14]. The TWD theory [15,16] have grown from the ideas of the rough set theory [17] to divide the universal set into positive, negative and boundary regions. Unlike the traditional two-way decision, the TWD incorporates the delay decision as an optional one. It is selected, if the cost of such delay is minimal [15]. It is of great importance in practice, besides taking a hard decision, to allow such "I do not know" option. There are several industrial applications of TWD in such data mining tasks, as visual feature extractions using deep neural networks [18], frequent item sets mining [19], attribute reduction [20], medical decision support systems [21], recommender systems [22] and software defect prediction [23]. However, the research of TWD in the classification problems for complex data has just begun [13]. Thus, in this paper we propose to examine the hierarchical representation of each segment using the methodology of granular computing [24,25]. The more detailed representation is explored only if the non-commitment option of TWD was chosen for the current representation.

The rest of the paper is organized as follows. In Sect. 2 we describe statistical speech recognition using an autoregression (AR) model [6,26]. In Sect. 3 we introduce the proposed classification algorithm based on sequential TWD. Section 4 contains experimental study of our approach in speech recognition for Russian language. Concluding comments are given in Sect. 5.

2 Conventional Classification of Piecewise-Stationary Speech Signals Using Statistical Approach

In this section we explore the task of isolated word recognition, which typically appears in, e.g., the voice control intelligent systems [27]. Let a vocabulary of $D > 1$ words/phrases be given. The dth word is usually specified by a sequence of phones $\{c_{d,1}, \ldots, c_{d,S_d}\}$. Here $c_{d,j} \in \{1, \ldots, C\}$ are the the class (phone) labels, and $S_d \geq 1$ is the transcription length of the dth word. It is required to assign the new utterance X to the closest word/phrase from the vocabulary. We focus on the speaker-dependent mode [6], i.e. the phonetic database of $R \geq C$ reference signals $\{\mathbf{x}_r\}, r \in \{1, \ldots, R\}$ with labels $c(r) \in \{1, \ldots, C\}$ of all phones of the current speaker should be available.

We use the typical assumption that the speech signal X can be represented as a piecewise stationary time-varying AR ergodic Gaussian process with zero mean [1,7,26]. To apply this model, the input utterance is divided into T fixed-size (20–30 ms) partially overlapped quasi-stationary frames $\{\mathbf{x}(t)\}, t \in \{1, \ldots, T\}$, where $\{\mathbf{x}(t)\}$ is a feature vector with the fixed dimension size. Next, each frame is assigned to one of C reference phones. It is known [28,29] that the maximal likelihood (ML) solution for testing hypothesis $W_c, c \in \{1, \ldots, C\}$ about covariance matrix of the Gaussian signal $\mathbf{x}(t)$ is achieved with the Kullback-Leibler (KL) minimum information discrimination principle [30]

$$c^*(\mathbf{x}(t)) = \underset{c(r), r \in \{1, \ldots, R\}}{\operatorname{argmin}} \rho_{KL}(\mathbf{x}(t), \mathbf{x}_r), \tag{1}$$

where the KL divergence between the zero-mean Gaussian distributions is computed as follows

$$\rho_{KL}(\mathbf{x}(t), \mathbf{x}_r) = \frac{1}{2} \ln \frac{\det(\Sigma_r)}{\det(\Sigma(t))} + \frac{1}{2} \operatorname{tr}(\Sigma(t)(\Sigma_r)^{-1}) - \frac{p}{2}.$$

Here $\Sigma(t)$ and Σ_r are the estimates of the covariance matrices of signals $\mathbf{x}(t)$ and $\mathbf{x}_r$, respectively, $\det(\Sigma)$ and $\operatorname{tr}(\Sigma)$ stand for the determinant and trace of the matrix Σ. This KL discrimination for the Gaussian model of the quasi-stationary speech signals can be computed as the Itakura-Saito distance [26,28] between power spectral densities (PSD) $G_{\mathbf{x}(t)}(f)$ and $G_r(f)$ of the input frame $\mathbf{x}(t)$ and $\mathbf{x}_r$:

$$\rho_{KL}(\mathbf{x}(t), \mathbf{x}_r) = \frac{2}{F} \sum_{f=1}^{F/2} \left(\frac{G_{\mathbf{x}(t)}(f)}{G_r(f)} - \ln \frac{G_{\mathbf{x}(t)}(f)}{G_r(f)} - 1 \right). \tag{2}$$

Here $f \in \{1, \ldots, F\}$, is the discrete frequency, and F is the sample rate (Hz). The PSDs in (2) can be estimated using the Levinson-Durbin algorithm and the Burg method [31]. The Itakura-Saito divergence between PSDs (2) is well known in speech processing due to its strong correlation with the subjective MOS (mean opinion score) estimate of speech closeness [6].

Finally, the obtained transcription $\{c^*(\mathbf{x}(1)), c^*(\mathbf{x}(2)), \ldots, c^*(\mathbf{x}(T))\}$ of the utterance X is dynamically aligned with the transcription of each word from the vocabulary to establish the temporary compliance between the sounds. Such alignment is implemented with the dynamic programming techniques, e.g., Dynamic Time Warping or the Viterbi algorithm in the HMM [6]. The decision can be made in favor to the closest word from the vocabulary in terms of the total conditional probability or, equivalently, the sum of distances (2).

The typical implementation of the described procedure includes the estimation of AR coefficients and the PSDs for each frame, matching with all phones (1), (2) and dynamic alignment with transcriptions of all words in the vocabulary. Thus, the runtime complexity of this algorithm is equal to

$O(F \cdot p \cdot T + R \cdot F \cdot T + T \cdot \sum_{d=1}^{D} S_d)$, where p is the order of AR model. The more is the count of frames T, the less is the recognition performance. Unfortunately, as it is written in introduction, the duration of every phone varies significantly even for the same speaker. Hence, the frame is usually chosen to be very small in order to contain only one quasi-stationary part of the speech signal. In the next section we propose to apply the TWD theory to speed-up the recognition procedure by using multi-scale representation of the speech segments.

3 Sequential Three-Way Decisions in Speech Recognition

3.1 Three-Way Decisions

Though speech recognition on the phonetic level at the present time is comparable in quality with the phoneme recognition by human [6], the variability sources (the noisy environment, children speech, foreign accents, speech rate, voice disease, etc.) usually lead to the misclassification errors [32]. Hence, in this paper we apply the TWD to represent each cth phone with three pair-wise disjoint regions (positive POS, negative NEG and boundary BND). These regions can be defined using the known asymptotic chi-squared distribution of the KL divergence between feature vectors of the same class [29,30]:

$$POS_{(\alpha,\beta)}(c) = \{\mathbf{x} \in \mathbf{X} | 2(n(\mathbf{x}) - p)\rho(\mathbf{x}, c) < \chi^2_{1-\alpha,p(p+1)/2}\}, \tag{3}$$

$$NEG_{(\alpha,\beta)}(c) = \{\mathbf{x} \in \mathbf{X} | 2(n(\mathbf{x}) - p)\rho(\mathbf{x}, c) \geq \chi^2_{1-\beta,p(p+1)/2}\}, \tag{4}$$

$$BND_{(\alpha,\beta)}(c) = \mathbf{X} - (POS_{(\alpha,\beta)}(c) \cup NEG_{(\alpha,\beta)}(c)), \tag{5}$$

where

$$\rho(\mathbf{x}, c) = \min_{r \in \{1,\ldots,R\}, c(r)=c} \rho_{KL}(\mathbf{x}, \mathbf{x}_r). \tag{6}$$

Here $\mathbf{X}$ is the universal set of the stationary speech signals, $n(\mathbf{x})$ is the count of samples in the signal $\mathbf{x}$, $\chi^2_{\alpha,p(p+1)/2}$ is the α-quantile of the chi-squared distribution with $p(p+1)/2$ degrees of freedom, $0 < \beta < \alpha < 1$ is the pair of thresholds, which define the type II and type I errors of the given utterance representing the cth phone. In this case the type I error is detected if the cth phoneme is not assigned to the positive region (3). The type II error takes place when the utterance from any other phoneme is not rejected (4).

3.2 Multi-class Three-Way Decisions

Though the described approach (3)-(5) can provide an additional robustness of speech recognition, it does not deal with the multi-scale nature of the speech signals [1]. To solve the issues with performance of traditional approach, we will use the multi-granulation approach [24,35] and describe the stationary utterance as a hierarchy of fragments. Namely, we obtain the largest piecewise quasi-stationary speech segments $X(s), s \in \{1, \ldots, S\}$

with the borders $(t_1(s), t_2(s)), 1 \leq t_1(s) < t_2(s) \leq T$ in observed utterance using an appropriate speech segmentation technique [1,11]. Here S is the count of extracted segments. Then, l speech parts of the same size are extracted at the lth granularity level, where the kth part $\mathbf{x}_k^{(l)}(s) = \left[\mathbf{x}\left(t_1(s) + \left\lfloor \frac{(k-1)\cdot(t_2(s)-t_1(s)+1)}{l} \right\rfloor\right), \ldots, \mathbf{x}\left(t_1(s) + \left\lceil \frac{k\cdot(t_2(s)-t_1(s)+1)}{l} \right\rceil\right)\right]$. Hence, only one part $\mathbf{x}_1^{(1)}(s) = X(s)$ of the sth segment is examined at the coarsest granularity level $l = 1$, and all $L = (t_2(s) - t_1(s) + 1)$ frames are processed at the finest granularity level.

According to the idea of sequential TWD [12], it is necessary to assign three decision regions at each granularity level. Though the concept of a phoneme is naturally mapped into TWD theory (3)-(5), speech recognition involves the choice of only one phoneme for each segment (1). Three basic options of acceptance, rejection and non-commitment are best interpreted in the binary classification task ($C = 2$) [15]. It includes three decision types: positive (accept the first class), negative (reject the first class and accept the second class), and boundary (delay the final decision and do not accept either first or second class). It cannot directly deal with *multi-class* problems ($C > 2$). This problem has been studied earlier in the context of multiple-category classification using decision-theoretic rough sets [34]. Lingras et al. [33] discussed the Bayesian decision procedure with C classes and specially constructed $2^C - 1$ cost functions. Liu et al. [37] proposed a two stages algorithm, in which, at first, the positive region is defined to make a decision of acceptance of any class, and the best candidate classification is chosen at the second stage using Bayesian discriminant analysis. Deng and Jia [36] derived positive, negative and boundary regions of each class from the cost matrix in classical cost-sensitive learning task.

However, in this paper we examine another enhancement of the idea of TWD for multi-class recognition, namely, $(C+1)$-way decisions, i.e., acceptance of any of C classes or delaying the decision process, in case of an unreliable recognition result [13]. In this case, it is necessary to define C positive regions $POS_{(\alpha,\beta)}^{(l)}(c)$ for each cth phone and one boundary region $BND_{(\alpha,\beta)}^{(l)}$ for delay option.

3.3 Proposed Approach

Let us aggregate the three regions of each phoneme (3)–(5) into such $(C+1)$-way decisions. The most obvious way is to assign an utterance $\mathbf{x}$ to the cth phone if this utterance is included into the positive region (3) of only this class:

$$POS_{(\alpha,\beta)}^{(l)}(c) = POS_{(\alpha,\beta)}(c) - \bigcup_{i \in \{1,\ldots,c-1,c+1,\ldots,C\}} POS_{(\alpha,\beta)}(i), \tag{7}$$

$$BND_{(\alpha,\beta)}^{(l)} = \mathbf{X} - \bigcup_{c=1}^{C} POS_{(\alpha,\beta)}^{(l)}(c). \tag{8}$$

It is not difficult to show, that the signal $\mathbf{x}$ is included into the positive region (7) of the nearest class $c^*(\mathbf{x})$ (1), only if

$$\begin{cases} 2(n(\mathbf{x})-p)\rho(\mathbf{x},c^*(\mathbf{x})) < \chi^2_{1-\alpha,p(p+1)/2} \\ 2(n(\mathbf{x})-p)\rho(\mathbf{x},c^*_2(\mathbf{x})) \geq \chi^2_{1-\alpha,p(p+1)/2} \end{cases}. \tag{9}$$

Here the second nearest neighbor class for the utterance $\mathbf{x}$ is denoted as

$$c^*_2(\mathbf{x}) = \underset{c\in\{1,\ldots,C\},c\neq c^*(\mathbf{x})}{\operatorname{argmin}} \rho(\mathbf{x},c). \tag{10}$$

However, in such definition of the positive regions the parameter α does not stand for the type I error anymore. As a matter of fact, the multiple-testing problem occurs in the multi-class classification, so appropriate correction should be used in the thresholds (9) [38]. If we would like to control the false discovery rate and accept the cth phone if only one hypothesis is accepted, the Benjamini-Hochberg test [39] with $(C-1)/C$ correction of type I error of the second hypothesis can be applied:

$$\begin{cases} 2(n(\mathbf{x})-p)\rho(\mathbf{x},c^*(\mathbf{x})) < \rho_1(\alpha) \\ 2(n(\mathbf{x})-p)\rho(\mathbf{x},c^*_2(\mathbf{x})) \geq \rho_2(\alpha) \end{cases}, \tag{11}$$

where the thresholds are defined as follows: $\rho_1(\alpha) = \chi^2_{1-\alpha,p(p+1)/2}$, $\rho_2(\alpha) = \chi^2_{1-\alpha(C-1)/C,p(p+1)/2}$. If condition (11) holds for all l parts at the lth granularity level, then the closest phones $c^*(\mathbf{x}^{(l)}_k(s))$ (1) are accepted as the final decisions. Otherwise, the delayed decision is chosen and the phoneme recognition problem is examined at a finer granulation level $l+1$ with more detailed information [12].

Unfortunately, the proposed procedure (11) can be hardly used in practice, because the distance between real utterances of the same phoneme is rather large and does not satisfy the theoretical chi-squared distribution with $p(p+1)/2$ degrees of freedom [29]. Hence, the first condition in (11) does not hold anymore. Thus, it is necessary to tune the thresholds ρ_1, ρ_2. However, in this paper we explore an alternative solution. Namely, the search termination condition (10) is modified by using the known probability distribution of the KL divergence between different hypothesis [30]. If the utterance $\mathbf{x}$ corresponds to the nearest neighbor phoneme $c^*(\mathbf{x})$, then the $2(n(\mathbf{x})-p)$-times distance $\rho(\mathbf{x},c^*_2(\mathbf{x}))$ is distributed as the non-central chi-squared distribution with $p(p+1)/2$ degrees of freedom and the non-centrality parameter proportional to the distance between phonemes $\rho(c^*(\mathbf{x}),c^*_2(\mathbf{x}))$ [27,40]. Thus, the ratio of the distances between the input signal and its second and first nearest neighbor has the non-central F-distribution $F(p(p+1)/2, p(p+1)/2; 2(n(\mathbf{x})-p))\rho(c^*(\mathbf{x}),c^*_2(\mathbf{x}))$. Hence, in this paper we will use the following positive region for acceptance of class c:

$$POS^{(l)}_{(\alpha,\beta)}(c) = \{\mathbf{x}\in\mathbf{X} | c = c^*(\mathbf{x}) \& \frac{\rho(\mathbf{x},c^*_2(\mathbf{x}))}{\rho(\mathbf{x},c^*(\mathbf{x}))} > \rho_{2/1}(\alpha)\}, \tag{12}$$

where a threshold $\rho_{2/1}(\alpha)$ is chosen from the α-quantile of the non-central F-distribution described above.

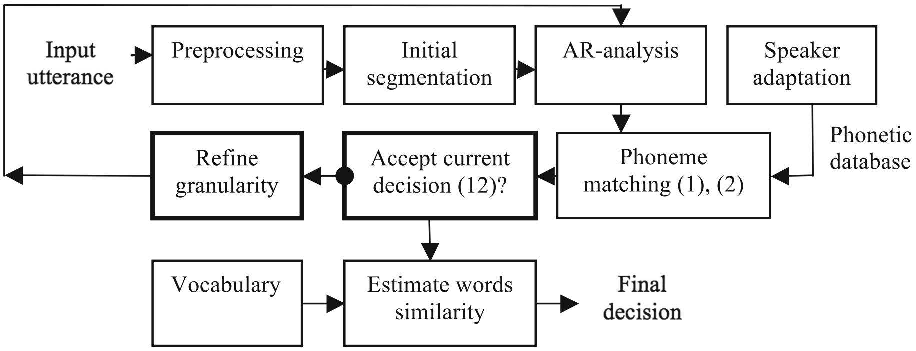

Fig. 1. Complete data flow of speech recognition using sequential three-way decisions and granular computing.

The complete data flow of the proposed recognition procedure using sequential TWD is shown in Fig. 1. At first, the input signal is preprocessed in order to decrease its variability, detect voice activity regions, etc. [6]. Next, the largest piecewise quasi-stationary speech segments are detected, and the coarsest approximation of the observed signal is analyzed. After that, each extracted segment is processed alternately. As we assume, that the scale of each part of the large segment $X(s)$ is identical, so the sequential analysis is terminated only when decisions are accepted for any speech part $\mathbf{x}_k^{(l)}(s)$. This procedure can be also implemented with the Benjamini-Hochberg correction of the type I error in (12). If it is possible to obtain a reliable solution $\mathbf{x}(s) \in POS^{(l)}_{(\alpha,\beta)}(c)$(12), the phoneme matching process (1), (2) is terminated and, as a result, the $c^*(\mathbf{x}(s))$ class label is assigned to this segment. Otherwise, its scale is refined, and the process is repeated for each part, until any of these parts are accepted (12). If the absence of acceptance decisions at all L levels for individual frames $\mathbf{x}(t)$, we can obtain the least unreliable level [13]. Finally, the estimated transcription of the refined segments can be processed using the dynamic programming techniques [6] in order to obtain the final decision of the speech recognition problem.

Let us demonstrate how the proposed procedure works in practice. In this example we consider rather simple task of Russian vowel recognition in a syllable "tro"(/t/ /r/ /oo/). Table 1 contains the KL distances (2) between $R = 6$ vowel phonemes and all segments in $L = 2$ hierarchical levels. The closest distance in each row is marked by bold. Here the vowel /aa/ is the nearest neighbor (1) of the signal $\mathbf{x}_1^{(1)}$ (see the first row in Table 1). Hence, the whole syllable ($l = 1$) is incorrectly classified. However, this decision cannot be accepted (12), because the distance to the second nearest neighbor /oo/ is quite close to the distance between $\mathbf{x}_1^{(1)}$ and the first nearest neighbor ($45.38/38.4 = 1.18$). Thus, according to sequential TWD scheme (Fig. 1) the granularity level should be refined, and the whole syllable is divided into $l = 2$ parts. Though the first part is still misclassified (second row in Table 1), this decision is still unacceptable as the distance ration in (12) is rather low (57.94/38.40 = 1.6). At the same time, the second part of the utterance is correctly recognized as the phone /oo/. This decision can be accepted (12), because the distance to the second nearest neighbor is rather large (46.02/5.83 = 7.89). As we know, that a syllable contains only one vowel, we can accept /oo/ phone as the final decision for the whole syllable. Thus, the proposed approach can be use to increase the recognition accuracy. In the next section we experimentally demonstrate that an additional refinement of the granularity level makes it possible to significantly decrease the decision making time.

Table 1. Computed distances (2) in the vowel recognition in the syllable /t/ /r/ /oo/

Level	/aa/	/ee/	/ii/	/oo/	/uu/	/y/
$l = 1$	**38.40**	85.38	270.80	45.38	113.33	99.28
$l = 2$	**36.17**	77.26	277.83	57.94	129.84	93.16
	333.39	303.94	198.47	**5.83**	46.02	460.33

4 Experimental Results

In this section the proposed approach (Fig. 1) in used in the isolated words recognition for Russian language. All tests are performed at a 4 core i7 laptop with 6 Gb RAM. Two vocabularies are used, namely, (1) the list of 1832 Russian cities with corresponding regions; and (2) the list of 1913 drugs. All speakers pronounced every word from all vocabularies twice in isolated syllable mode to simplify the recognition procedure [27,40]. In such mode every vowel in the syllable is made stressed, thus, it is recognized quite stably. The part of speech data suitable to reproduce our experiments is available for free download[1]. In the configuration mode, each speaker clearly spoke ten vowels of the Russian language (/aa/, /ja/, /ee/, /je/, /oo/, /jo/, /ii/, /y/, /uu/, /ju/) in isolated mode [41]. The following parameters are chosen: sampling frequency $F = 8$ kHz, AR-model order $p = 20$. The sampling rate was set on telephone level, because we carried out this experiment with our special software [27,42], which was mainly developed for application in remote voice control systems.

The closed sounds /aa/, /ja/, /ee/, /je/, /oo/, /jo/, /ii/, /y/, /uu/, /ju/ are united into $C = 5$ clusters [6]. Observed utterances are divided into 30 ms frames with 10 ms overlap. The syllables in the test signals are extracted with the amplitude detector and the vowels are recognized in each syllable by the simple voting [40] based on the results obtained using vowel recognition. The latter is implemented using either proposed sequential TWD procedure with termination condition (12), or traditional techniques: (1) recognition (1), (2) of low-scale frames with identical size; (2) distance thresholding (11); and (2) the

[1] https://sites.google.com/site/andreyvsavchenko/SpeechDataIsolatedSyllables.zip.

state-of-the-art recognition of vowels in each syllable using the DNN from the Kaldi framework [43] trained with the Voxforge corpus. We added an artificially generated white noise to each test utterance using the following procedure. At first, the signal-to-noise ratio (SNR) is fixed. Next, the pauses are detected in each utterance using simple energy thresholding, and the standard deviation of the remaining part with high energy is estimated. Finally, these standard deviation was corrected using given SNR, and uncorrelated normal random numbers with zero mean and the resulted standard deviation was added to each value of the speech signal.

Except the KL divergence (2), its symmetric version (COSH distance [2,28]) is implemented:

$$\rho_{COSH}(\mathbf{x}(t), \mathbf{x}_r) = \frac{1}{F} \sum_{f=1}^{F/2} \frac{(G_{\mathbf{x}(t)}(f) - G_r(f))^2}{G_{\mathbf{x}(t)}(f) G_r(f)}. \tag{13}$$

The thresholds in (11), (12) for each discrimination type are tuned experimentally using the small validation set of 5 vowels per phone class[2]. Namely, we compute the pairwise distances between all utterances from this validation set $\mathbf{X}_{val}$. If type I error rate is fixed $\alpha = const$, then $\rho_{2/1}(\alpha)$ is evaluated as a $(1-\alpha)$-quantile of the ratio of these distances

$$\left\{ \frac{\min\limits_{\mathbf{x}_r \in \mathbf{X}_{val}, c(\mathbf{x}_r) \neq c(\mathbf{x})} \rho(\mathbf{x}, \mathbf{x}_r)}{\min\limits_{\mathbf{x}_r \in \mathbf{X}_{val}, \mathbf{x}_r \neq \mathbf{x}} \rho(\mathbf{x}, \mathbf{x}_r)} \middle| \mathbf{x} \in \mathbf{X}_{val} \right\}.$$

Similar procedure is applied to estimate thresholds in (11) [5]. The dependence of the words recognition accuracy on the SNR is shown in Tables 2 and 3 for cities and drugs vocabularies, respectively. The average time to recognize one testing phrase is shown in Figs. 2 and 3.

Table 2. Dependence of error rate (%) on SNR (dB), cities vocabulary

Distance	Method	25 dB	20 dB	15 dB	10 dB	5 dB	0 dB
	DNN	6.3	7.9	10.2	18.9	30.6	34.1
	Conventional approach (1)	7	7.9	8.8	14.5	31.6	38.2
KL divergence	Distance thresholding (11)	7.5	8.8	10.5	18.9	30.1	36.1
	Proposed approach (12)	6.3	7.3	9.6	17.3	31.1	37.1
	Conventional approach (1)	3.9	4.2	4.1	9.8	23.7	28.1
COSH distance	Distance thresholding (11)	3.1	3.9	4.7	10.8	26	32.7
	Proposed approach (12)	3.9	4.7	5.3	11.8	26.9	33.9

Though the state-of-the-art DNN does not use speaker adaptation, its accuracy of vowel recognition is comparable to the nearest neighbor search (1), which

[2] https://sites.google.com/site/andreyvsavchenko/ValidationDataVowels.zip.

Table 3. Dependence of error rate (%) on SNR (dB), drugs vocabulary

Distance	Method	25 dB	20 dB	15 dB	10 dB	5 dB	0 dB
KL divergence	DNN	9.9	10.6	11.4	13.9	18.4	23.2
	Conventional approach (1)	3.1	5.4	8.1	8.3	15.9	20.4
	Distance thresholding (11)	4.1	6.6	8.7	8.7	17	20.3
	Proposed approach (12)	3.9	6.6	8.3	8.6	15.9	19.9
COSH distance	Conventional approach (1)	5.6	6.6	6.8	6.8	14.3	17.4
	Distance thresholding (11)	3.5	4.3	7.5	7.9	14.1	18.6
	Proposed approach (12)	2.9	3.7	7.5	8.1	14.1	17.4

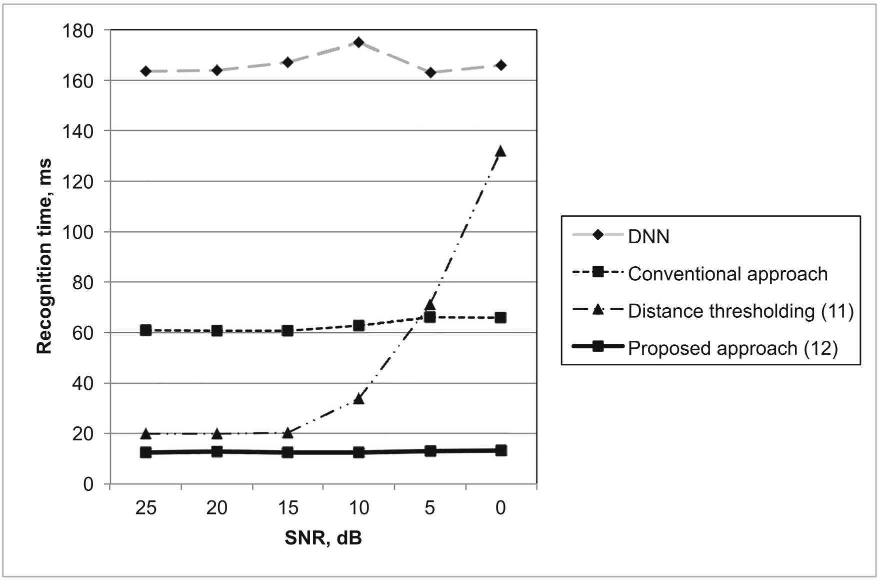

Fig. 2. Experimental results, cities vocabulary.

is implemented in other examined techniques. However, the DNN's performance is inappropriate: it is 2–10 times slower than all other methods. McNemar's test [44] with 0.95 confidence verified that the COSH distance is more accurate in most cases, than the KL divergence. This result supports our statement about superiority of the distances based on the homogeneity testing in audio and visual recognition tasks [2]. The obvious implementation of sequential TWD (11) is inefficient in the case of high noise levels, because the thresholds in (11) cannot be reliably estimated for huge variations in speech signals. Finally, the proposed approach (Fig. 1) allows to increasing the recognition performance. Our implementation of sequential TWD is 12–14 times faster that the DNN and 4–5 times faster than the conventional approach with matching of the fine-grained frames (1). McNemar's test verified that this improvement of performance is significant

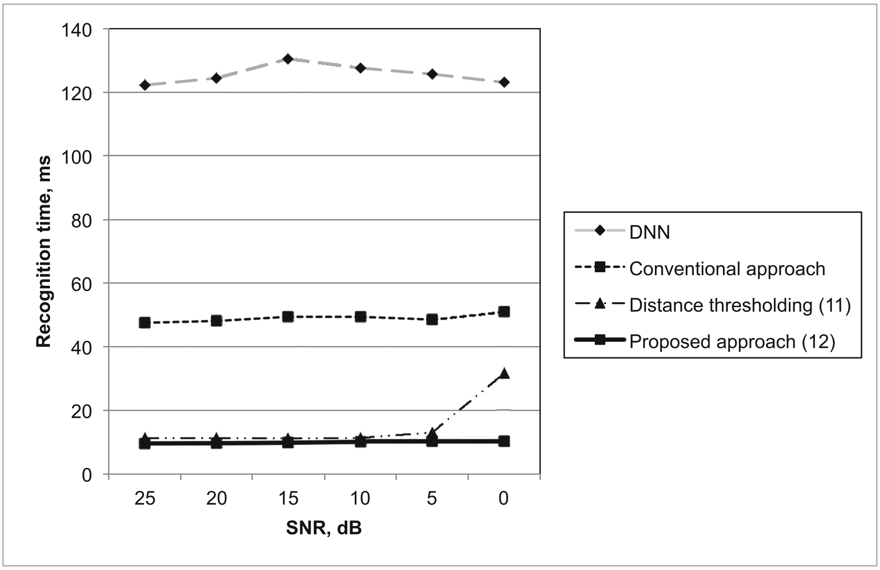

Fig. 3. Experimental results, drugs vocabulary.

in all cases except the experiment with drugs vocabulary (Fig. 3), in which classification speed of both (11) and (12) is similar for low level of noise ($SNR > 5$). Moreover, our approach leads to the most accurate decisions (for a fixed dissimilarity measure) in all cases except the recognition of drugs (Table 3) with the KL-divergence (2) and low noise level. However, these differences in error rates are mostly not statistically significant.

5 Conclusion

To sum it up, this article introduced an efficient implementation (1), (6), (10), (12) of sequential three-way decisions in multi-class recognition of piecewise stationary signals. It was demonstrated how to define the granularity levels in quasi-stationary parts of the signal, so that the count of the coarse-grained granules is usually rather low. As a result, the new observation can be classified very fast. The acceptance region (12) was defined using the theory of multiple comparisons and contains only computing the KL divergence. Hence, our method can be applied with an arbitrary distance by tuning the threshold $\rho_{2/1}$. The experimental study demonstrated the potential of our procedure (Fig. 1) to significantly speed-up speech recognition when compared with conventional algorithms (Figs. 2 and 3). Thus, it is possible to conclude that the proposed technique makes it possible to build a reliable speech recognition module, which is suitable for implementing, e.g., a voice control intelligent system with fast speaker adaptation [27].

As a matter of fact, our experiments are reported on own speech data with requirement of isolated syllable pronunciation. Thus, our results are not directly comparable with other ASR methods. Hence, the further research of the proposed method can be continued in the following directions. First, it should be applied in continuous speech recognition, in which only the last granularity level is analyzed with the computationally expensive state-of-the-art procedures (HMMs with GMMs/DNNs or LSTMs) [6,9,10]. Second possible direction is the application of our method with non-stationary signal classification tasks [4].

Acknowledgements.. The work was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017 (grant No 17-05-0007) and is supported by the Russian Academic Excellence Project “5–100” and Russian Federation President grant no. MD-306.2017.9.

References

1. Tyagi, V., Bourlard, H., Wellekens, C.: On variable-scale piecewise stationary spectral analysis of speech signals for ASR. Speech Commun. **48**, 1182–1191 (2006)
2. Savchenko, A.V., Belova, N.S.: Statistical testing of segment homogeneity in classification of piecewise-regular objects. Int. J. Appl. Math. Comput. Sci. **25**, 915–925 (2015)
3. Huang, K., Aviyente, S.: Sparse representation for signal classification. In: Advances of Neural Information Processing Systems (NIPS), pp. 609–616. MIT Press (2006)
4. Khan, M.R., Padhi, S.K., Sahu, B.N., Behera, S.: Non stationary signal analysis and classification using FTT transform and naive bayes classifier. In: IEEE Power, Communication and Information Technology Conference (PCITC), pp. 967–972. IEEE Press (2015)
5. Savchenko, A.V.: Search Techniques in Intelligent Classification Systems. Springer International Publishing, New York (2016)
6. Benesty, J., Sondhi, M.M., Huang, Y.: Springer Handbook of Speech Processing. Springer, Berlin (2008)
7. Peebles, P.Z., Read, J., Read, P.: Probability, Random Variables, and Random Signal Principles. McGraw-Hill, New York (2001)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press, Cambridge (2016)
9. Yu, D., Deng, L.: Automatic Speech Recognition: A Deep Learning Approach. Springer, New York (2014)
10. Sak, H., Senior, A.W., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Interspeech, pp. 338–342 (2014)
11. Stan, A., Mamiya, Y., Yamagishi, J., Bell, P., Watts, O., Clark, R.A., King, S.: ALISA: an automatic lightly supervised speech segmentation and alignment tool. Comput. Speech Lang. **35**, 116–133 (2016)
12. Yao, Y.Y.: Granular computing and sequential three-way decisions. In: Lingras, P., Wolski, M., Cornelis, C., Mitra, S., Wasilewski, P. (eds.) RSKT 2013. LNCS (LNAI), vol. 8171, pp. 16–27. Springer, Heidelberg (2013)

13. Savchenko, A.V.: Fast multi-class recognition of piecewise regular objects based on sequential three-way decisions and granular computing. Knowl.-Based Syst. **91**, 252–262 (2016)
14. Li, H., Zhang, L., Huang, B., Zhou, X.: Sequential three-way decision and granulation for cost-sensitive face recognition. Knowl.-Based Syst. **91**, 241–251 (2016)
15. Yao, Y.: Three-way decisions with probabilistic rough sets. Inf. Sci. **180**, 341–353 (2010)
16. Yao, Y.: Interval sets and three-way concept analysis in incomplete contexts. Int. J. Mach. Learn. Cybern. **8**(1), 1–18 (2017)
17. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Norwell, MA, USA (1992)
18. Li, H., Zhang, L., Zhou, X., Huang, B.: Cost-sensitive sequential three-way decision modeling using a deep neural network. Int. J. Approx. Reason. **85**, 68–78 (2017)
19. Li, Y., Zhang, Z.H., Chen, W.B., Min, F.: TDUP: an approach to incremental mining of frequent itemsets with three-way-decision pattern updating. Int. J. Mach. Learn. Cybern. **8**(2), 441–453 (2017)
20. Ren, R., Wei, L.: The attribute reductions of three-way concept lattices. Knowl.-Based Syst. **99**, 92–102 (2016)
21. Yao, J., Azam, N.: Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets. IEEE Trans. Fuzzy Syst. **23**(1), 3–15 (2015)
22. Zhang, H.R., Min, F., Shi, B.: Regression-based three-way recommendation. Inf. Sci. **378**, 444–461 (2017)
23. Li, W., Huang, Z., Li, Q.: Three-way decisions based software defect prediction. Knowl.-Based Syst. **91**, 263–274 (2016)
24. Pedrycz, W.: Granular Computing: Analysis and Design of Intelligent Systems. CRC Press, Boca Raton (2013)
25. Wang, X., Pedrycz, W., Gacek, A., Liu, X.: From numeric data to information granules: a design through clustering and the principle of justifiable granularity. Knowl.-Based Syst. **101**, 100–113 (2016)
26. Itakura, F.: Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust. Speech Signal Process. **23**, 67–72 (1975)
27. Savchenko, A.V., Savchenko, L.V.: Towards the creation of reliable voice control system based on a fuzzy approach. Pattern Recognit. Lett. **65**, 145–151 (2015)
28. Gray, R.M., Buzo, A., Gray, J.A., Matsuyama, Y.: Distortion Measures for Speech Processing. IEEE Trans. Acoust. Speech Signal Process. **28**, 367–376 (1980)
29. Savchenko, V.V., Savchenko, A.V.: Information-theoretic analysis of efficiency of the phonetic encoding-decoding method in automatic speech recognition. J. Commun. Technol. Electron. **61**, 430–435 (2016)
30. Kullback, S.: Information Theory and Statistics. Dover Publications, New York (1997)
31. Marple, S.L.: Digital Spectral Analysis: With Applications. Prentice Hall, Upper Saddle River (1987)
32. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C.: Automatic speech recognition and speech variability: a review. Speech Commun. **49**, 763–786 (2007)
33. Lingras, P., Chen, M., Miao, D.: Rough multi-category decision theoretic framework. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS, vol. 5009, pp. 676–683. Springer, Heidelberg (2008). doi:10.1007/978-3-540-79721-0_90

34. Zhou, B.: Multi-class decision-theoretic rough sets. Int. J. Approx. Reason. **55**(1), 211–224 (2014)
35. Ju, H.R., Li, H.X., Yang, X.B., Zhou, X.Z.: Cost-sensitive rough set: a multi-granulation approach. Knowl.-Based Syst. **123**, 137–153 (2017)
36. Deng, G., Jia, X.: A decision-theoretic rough set approach to multi-class cost-sensitive classification. In: Flores, V., et al. (eds.) IJCRS 2016. LNCS, vol. 9920, pp. 250–260. Springer, Cham (2016). doi:10.1007/978-3-319-47160-0_23
37. Liu, D., Li, T., Li, H.: A multiple-category classification approach with decision-theoretic rough sets. Fundam. Inform. **115**(2–3), 173–188 (2012)
38. Hochberg, Y., Tamhane, A.C.: Multiple Comparison Procedures. Wiley, Hoboken (2009)
39. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. Series B (Methodol.) **57**(1), 289–300 (1995)
40. Savchenko, A.V., Savchenko, L.V.: Classification of a sequence of objects with the fuzzy decoding method. In: Cornelis, C., Kryszkiewicz, M., Ślęzak, D., Ruiz, E.M., Bello, R., Shang, L. (eds.) RSCTC 2014. LNCS, vol. 8536, pp. 309–318. Springer, Cham (2014). doi:10.1007/978-3-319-08644-6_32
41. Savchenko, A.V.: Semi-automated speaker adaptation: how to control the quality of adaptation? In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2014. LNCS, vol. 8509, pp. 638–646. Springer, Cham (2014). doi:10.1007/978-3-319-07998-1_73
42. Savchenko, A.V.: Phonetic words decoding software in the problem of Russian speech recognition. Autom. Remote Control **74**(7), 1225–1232 (2013)
43. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
44. Gillick, L., Cox, S.: Some statistical issues in the comparison of speech recognition algorithms. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 532–535 (1989)