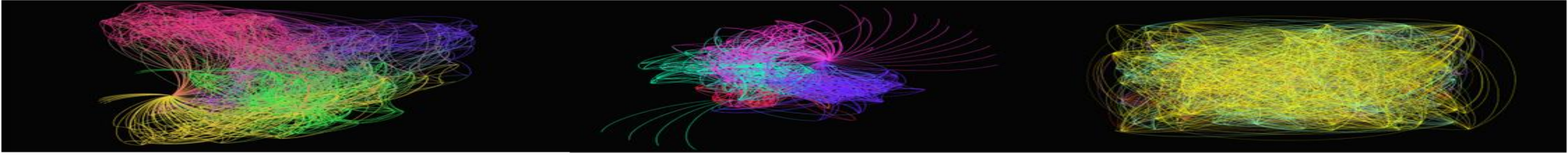


Workshop on Clustering and Search techniques in large scale networks

National Research University
Higher School of Economics
Nizhny Novgorod

Statistical classification of a sequence of objects based on a fuzzy approach

Andrey Savchenko
LATNA
PhD, Associate Professor

2015

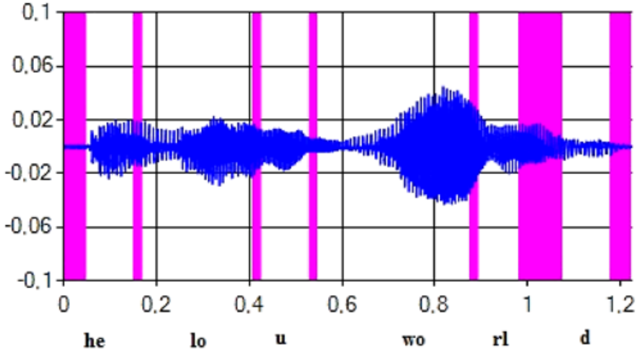
The problem of recognition of a set of objects

What for?

Let the input sequence $\{X(t)\}$ of $T>1$ frames be specified. It is assumed that different *observations* of only **one** object are presented in this sequence. The problem is to **assign** this sequence to one of $R>1$ **classes** specified by the **reference instances** $\{X_r\}$. This problem usually appears as a part of complex object or speech recognition systems.

Examples

Object $X(t)$ is a feature vector of one speech frame (in a phoneme recognition problem)



Object $X(t)$ is a single image (in still-to-still video-based face recognition problem)

Key idea

Improve the quality of SV by defining each class as a fuzzy set of all available instances

And now we introduce the agenda of our talk

1. State-of-the-art: Simple Voting (SV) and statistical approach

2. Fuzzy Decoding (FD) Method

3. Experimental results in phoneme and speech recognition

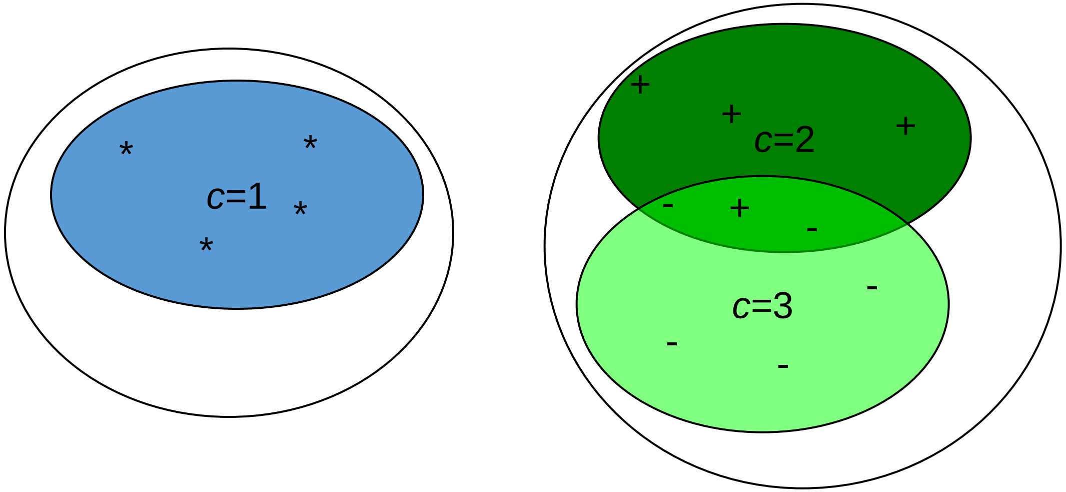
4. Concluding comments

Conventional approach

$I \geq 1$ reference instances are given for each class r.

Centroid-based classification Rocchio algorithm): Centroid of the r-th class:

$$X_r = \underset{\mathbf{x}_{r,k}, k \in \{1,\dots,I\}}{\arg\min} \sum_{i=1}^{I} \rho(\mathbf{x}_{r,k}, \mathbf{x}_{r,i})$$



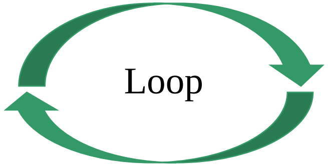
Disadvantage

The mathematical models of each class are independent. No information about classes similarities. Sometimes closed classes are united into one cluster

Simple voting method summary

Statistical pattern recognition + SV

Assumption	Bayesian decision for each frame	Aggregation by simple voting
Objects in each class are *identically distributed* and all distributions are of **multivariate exponential type** $f_{\boldsymbol{\theta}}(X)$ generated by the fixed (for all classes) function $f_0(X)$ and the parameter vector $\boldsymbol{\theta}$. Its ***unbiased*** consistent estimation: $\hat{\boldsymbol{\theta}}(X_r)$	Classification of observation $X(t)$ by the nearest neighbor rule $v(t) = \arg\min\limits_{r=\overline{1,R}} \hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}(X_r)}; X(t)\right)$ with the Kullback-Leibler (KL) divergence $\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}(X_r)}; X(t)\right) = \int f_{\hat{\boldsymbol{\theta}}(X(t))}(X) \cdot \ln \frac{f_{\hat{\boldsymbol{\theta}}(X(t))}(X)}{f_{\hat{\boldsymbol{\theta}}(X_r)}(X)} dX$	Solution is made in favor of the most frequent class $r^* = \arg\max\limits_{r=\overline{1,R}} \mu_r$ $\mu_r = \sum\limits_{t=1}^{T} \delta(v(t) - r)$ δ() - discrete Dirac delta function



KL divergence between instances of 2 classes characterize information to distinguish objects from these classes

Exponential family: Overview

Definition

Distribution $f_{\theta}(X)$ generated by function $f_0(X)$ with parameter vector $\boldsymbol{\theta}$

$$f_{\boldsymbol{\theta}}(X) = \exp\left(\boldsymbol{\tau}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}}(X)\right) \cdot f_0(X) / M(\boldsymbol{\tau}),$$

$$M(\boldsymbol{\tau}) = \int \exp\left(\boldsymbol{\tau}(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{\theta}}(X)\right) \cdot f_0(X) dX$$

If the parameter estimation $\hat{\boldsymbol{\theta}}(X)$ is unbiased, normalizing function (K-dimensional parameter vector) is defined by equation

$$\int \hat{\boldsymbol{\theta}}(X) \cdot f_{\boldsymbol{\theta}}(X) dX \equiv \frac{d}{d\boldsymbol{\tau}} \ln M(\boldsymbol{\tau}) = \boldsymbol{\theta}$$

Sample distributions

1. *Binomial*. *x*- number of successes in *n* yes/no experiments with success probability *p*, *X* - random variable

$$\boldsymbol{\theta} = [n \cdot p], \hat{\boldsymbol{\theta}}(X) = [x], \boldsymbol{\tau} = [\tau], f_0(X) = C_n^x$$

$$M(\boldsymbol{\tau}) = (1 + \exp(\tau))^n, \frac{d}{d\boldsymbol{\tau}} \ln M(\boldsymbol{\tau}) = \frac{n \exp(\tau)}{1 + \exp(\tau)} = np$$

$$\tau = \ln \frac{p}{1-p}, \boxed{f(x) = C_n^x \cdot p^x \cdot (1-p)^{n-x}}$$

2. *Normal* N(0;σ^2). Random sample is given.

$$\boldsymbol{\theta} = \left[\sigma^2\right], \hat{\boldsymbol{\theta}}(X) = \left[s^2\right], s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i)^2, f_0(X) = 1$$

$$M(\boldsymbol{\tau}) == \left(-\frac{\pi n}{\tau}\right)^{n/2}, \frac{d}{d\boldsymbol{\tau}} \ln M(\boldsymbol{\tau}) = -\frac{n}{2\tau} = \sigma^2$$

$$\tau = -\frac{n}{2\sigma^2}, \boxed{f(x_1, \ldots, x_n) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \cdot e^{-\frac{n \cdot s^2}{2\sigma^2}}}$$

3. Most common distributions: normal, multinomial, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson

KL discrimination

1. *Binomial*.

$$\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}(X_r)}; X\right) =$$

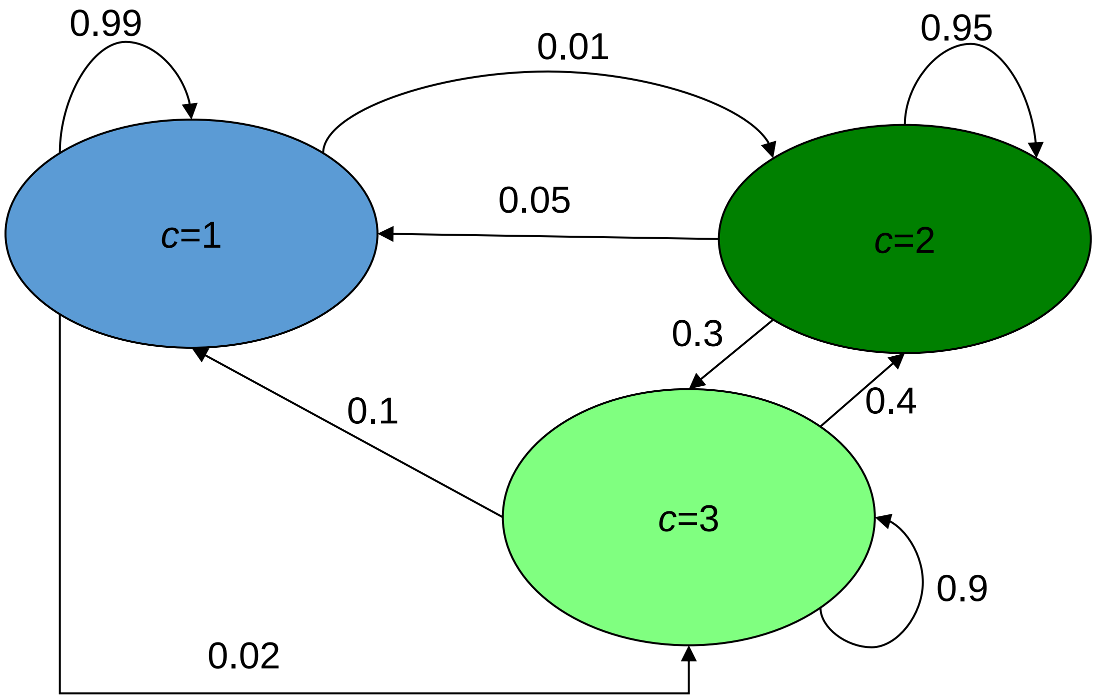
$$= x \ln \frac{x/n}{x_r/n_r} + (n-x) \ln \frac{(1-x/n)}{(1-x_r/n_r)}$$

2. *Normal*.

$$\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}(X_r)}; X\right) = \frac{n}{2}\left(\ln \frac{\hat{s}_r^2}{\hat{s}^2} - 1 + \frac{\hat{s}^2}{\hat{s}_r^2}\right)$$

Our purpose

Improve conventional approach by using the known distances between classes



Fuzzy Sets

L. Zadeh *Fuzzy sets // Information and Control*. 1965

Universal set $\mathbf{X} = \{x_1, \dots, x_N\}$

Fuzzy set $A = \left\{ (x_i, \mu_i^{(A)}) \middle| x_i \in \mathbf{X} \right\}$

Degree of membership $\mu_i^{(A)} \in [0;1]$ Compare with an ordinary (crisp) set: $\mu_i^{(A)} \in \{0;1\}$

Example. Closed to 5 numbers $\{(3;0.3),(4;0.7),(5;1),(6;0.7),(7;0.3)\} \equiv \frac{3}{0.3} + \frac{4}{0.7} + \frac{5}{1} + \frac{6}{0.7} + \frac{7}{0.3}$

Operations

Fuzzy union $A \cup B = \left\{ (x_i, \max(\mu_i^{(A)}, \mu_i^{(B)}) \middle| x_i \in \mathbf{X} \right\}$

Fuzzy intersection $A \cap B = \left\{ (x_i, \min(\mu_i^{(A)}, \mu_i^{(B)}) \middle| x_i \in \mathbf{X} \right\}$

Conditional probabilities estimation

1 **Confusion probability** of marking object from *j*-th class as r-th class (i.e., the distance between the object from *j*-th class and X_r is minimal).
If X is the object from *j*-th class then

$$2\cdot\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}}(X_r); X\right)$$

is asymptotically distributed as the non-central χ^2 with (K-1) degrees of freedom and noncentrality parameter:

$$2\cdot\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}}(X_r); X_j\right)$$

Confusion probability is estimated with the known distribution of independent minimum normal variables

$$P\left(X_r|X_j\right)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}\exp\left(-t^2/2\right)dt\prod_{\substack{i=1\\ i\neq r}}^{R}\left(1-\Phi\left(\frac{t\cdot\sqrt{8\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}}(X_r); X_j\right)+K-1}+2\left(\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}}(X_r); X_j\right)-\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}}(X_i); X_j\right)\right)}{\sqrt{8\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}}(X_i); X_j\right)+K-1}}\right)\right)$$

2 **Posterior probability** of $X(t)$ is estimated from the known relationship of the KL divergence and the maximal likelihood

$$P\left(X_r|X(t)\right)=\frac{\exp\left(-\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}}(X_r); X(t)\right)\right)}{\sum\limits_{i=1}^{R}\exp\left(-\hat{I}\left(*: f_{\hat{\boldsymbol{\theta}}}(X_i); X(t)\right)\right)}$$

Fuzzy Decoding Method

MAIN PROPOSAL: each *j*-th class is represented not only by an instance X_j, but by a **fuzzy set**

1 Each reference $\mathbf{x}_{r;i}$ is associated with the fuzzy set:

$$\left\{\left(X_r, \mu_r^{(j)}\right)\right\}, \mu_r^{(j)} = P\left(\mathbf{x}_{r;i} \middle| X_j\right)$$

2 Each class is associated with the fuzzy set by using the **fuzzy union**:

$$\left\{\left(X_r, \mu_r^{(j)}\right)\right\}, \mu_r^{(j)} = \max_{i \in \{1,\ldots,I\}} P\left(\mathbf{x}_{r;i} \middle| X_j\right)$$

3 Each *t*-th frame is associated with fuzzy set of posterior probabilities

$$\left\{\left(X_r, \mu_r(X(t))\right)\right\}, \mu_r(X(t)) = P\left(X_r \middle| X(t)\right)$$

4 To verify the correctness of the nearest neighbor class $v(t)$, perform the **fuzzy intersection**:

$$\mu(r;t) = \min\left(\mu_r^{(v(t))}, \mu_r(X(t))\right)$$

It is known (Kullback, 1997) that, if $X(t)$ belongs to the same class as the reference X_γ and if $\gamma = v(t)$, then $\mu(v(t);t) \approx 1$.
In case of recognition error $\mu(v(t);t) << 1$.

Fuzzy union preserves the value of final degree if any of references $\mathbf{x}_{j;i}$ is closed to the frame $X(t)$.

Fuzzy Decoding (FD) Method for an arbitrary distance

Described estimations include only calculation of the KL divergence. It can be replaced to an arbitrary distance with an appropriate smoothing factor α

1 PRELIMINARY STEP

1.1 Associate *j*-th class with fuzzy set of classes confusions

$$\boxed{\mu_r^{(j)} = \max_{i\in\{1,...,I\}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-t^2/2\right) \prod_{i=1, i\neq r}^{R} \left(1-\Phi\left(\frac{t\cdot\sqrt{8\alpha\cdot\rho(X_j,\mathbf{x}_{r;i})+K-1}+2\alpha\cdot\left(\rho(X_j,\mathbf{x}_{r;i})-\rho(X_j,\mathbf{x}_{r;i})\right)}{\sqrt{8\alpha\cdot\rho(X_j,X_i)+K-1}}\right)\right)dt}$$

2 RECOGNITION PROCEDURE

2.1 For each *t*-th frame

2.1.1 Associate *t*-th frame with fuzzy set of posterior probabilities

$$\boxed{\mu_r(X(t)) = \frac{\exp\left(-\alpha\cdot\rho\left(X(t),X_r\right)\right)}{\sum_{i=1}^{R}\exp\left(-\alpha\cdot\rho\left(X(t),X_i\right)\right)}}$$

2.1.2 Obtain the nearest neighbor $\nu(t) = \arg\min_{r=\overline{1,R}} \rho\left(X(t),X_r\right)$

2.1.3 Perform the fuzzy intersection $\mu(r;t) = \min\left(\mu_r^{(\nu(t))}, \mu_r(X(t))\right)$

2.2 Aggregate intersections for all frames $\mu_r = \frac{1}{T}\sum_{t=1}^{T}\mu(r;t)$

2.3 Final solution is made in favor of the class $r^* = \arg\max_{r=\overline{1,R}} \mu_r$

Here is exactly how does our FD method works. Example of recognition of the phone /y/ (/ы/) in a syllable “tj” (“ты”)

Simple Voting results (several phones are united into one cluster)

	Phone				
	/u/	/ju/	/je/	/ee/	/y/
Frequency rate	0.8	0.8	0.2	0.2	0

Processing of one frame in the FD

	Phone			
	/u/	/ju/	/je/	/y/
$\mu_r(X(t))$	0.123	0.1932	0.0858	0.1027
$\mu_r^{(v(t))}$	0.107	0.1102	0.1052	0.1072
$\mu(r;t)$	0.107	0.1102	0.0858	0.1027

Fuzzy Decoding method results

	Phone			
	/u/	/ju/	/je/	/y/
μ_r	0.22	***0.34***	0.18	0.26

Better recognition results though further clarification (lexical, semantic, etc.) is needed

No need to unite closed phonemes (e.g., /u/ and /ju/) into one cluster

Experiments. Synthetic dataset

Recognition of the normally distributed random signals

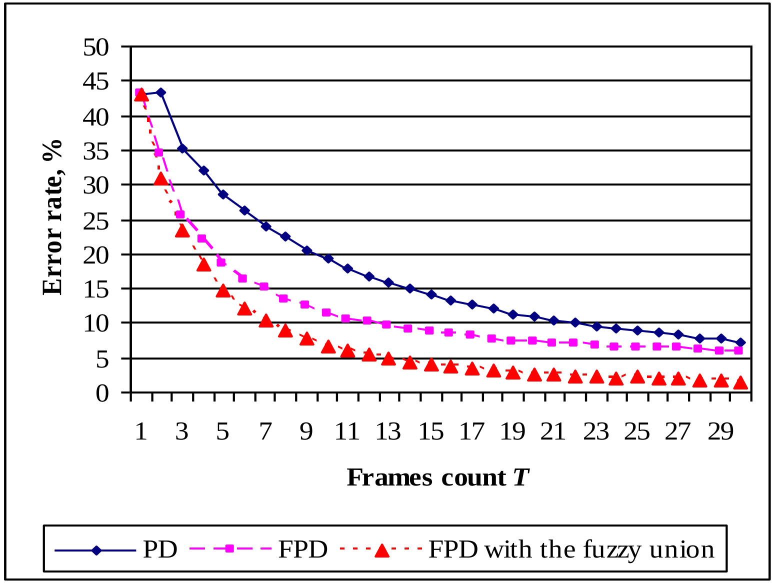
Covariance matrix size 2x2 Mean: (0,0)

Number of classes *R* 10 with correlation coefficients: 0.1 0.2 0.3 … 0.8 0.9

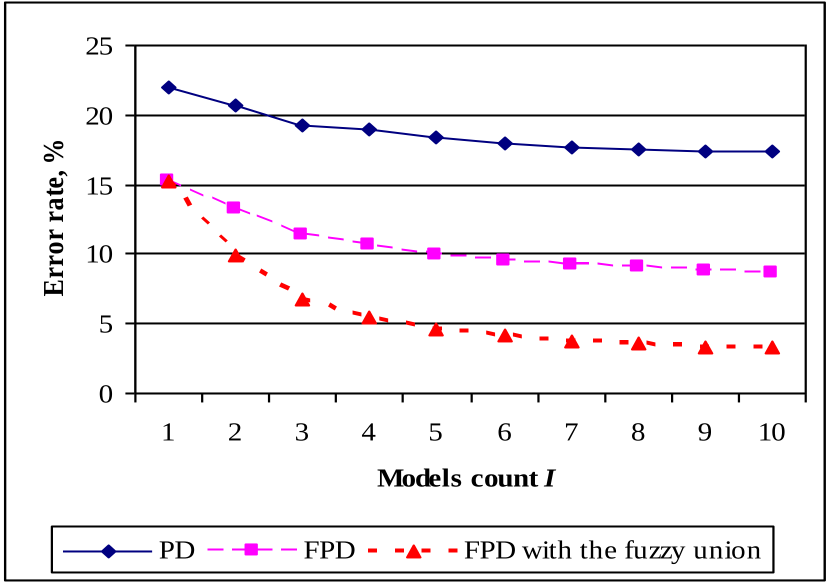
Reference instance is generated by adding random variable *N*(0;0.03) to the correlation coefficient of the class

Test signal is generated by adding random variable *N*(0;0.07) to the correlation coefficient of the class

Dependence of the error rate (%) on the number of frames *T* (*I*=3 references per class)



Dependence of the error rate (%) on the number of instances *I* (*T*=10 frames)



Experiments. Speech recognition

State-of-the-art similarity measures and speech features:

1 12 MFCCs (Mel-Frequency Cepstral Coefficients) features + their first derivatives (totally, *K=24* parameters) compared with the Euclidean distance

2 Autoregression (AR) estimates of the speech signal PSD (Power Spectral Densities) obtained with Levinson-Durbin algorithm and Burg method

2.1 Itakura-Saito (IS) divergence equivalent with a constant factor to the KL discrimination for Gaussian signals

$$\rho_{IS}(X_1,X_2)=\frac{1}{F}\sum_{f=1}^{F}\left(\frac{G_1(f)}{G_2(f)}-\ln\frac{G_1(f)}{G_2(f)}-1\right)$$

where $G_1(f)$, $G_2(f)$ – PSDs of signals X_1, X_2

2.2 Spectral distortion (SD) which is the known equivalent to the linear prediction coding cepstral coefficients' comparison in Euclidean space

$$\rho_{SD}(X_1,X_2)=\frac{1}{F}\sum_{f=1}^{F}\left(\ln G_1(f)-\ln G_2(f)\right)^2$$

The experimental results were obtained with the following parameters

Number of speakers 5 where: 3 male 2 female

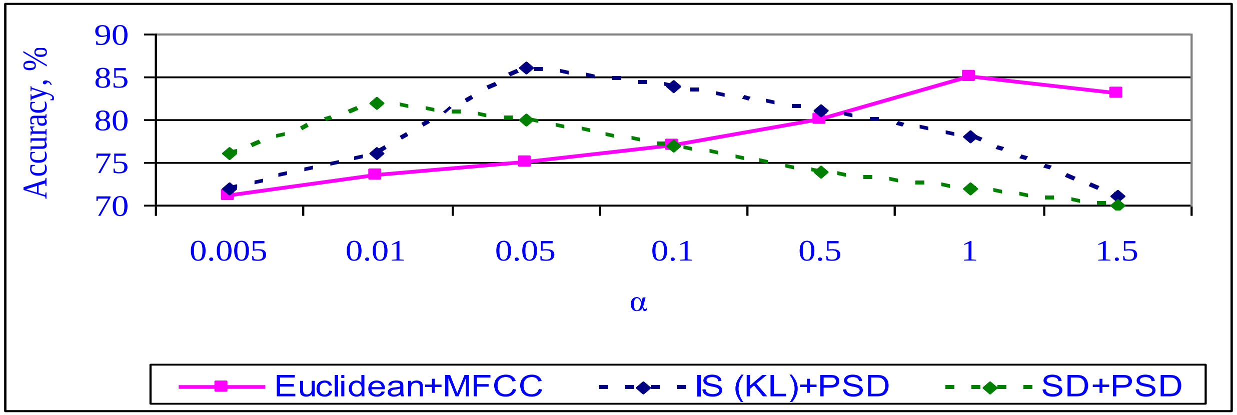
Number of model vowels *R* 10 with: 5 clusters of phonemes for SV Speaker-dependent mode

Sampling frequency 8 kHz AR-order: 20 Frame duration: 45 ms. Overlap: 30 ms.

Distance smoothing factor α: 0.005 0.001 0.05 0.1 0.5 1 1.5

Experimental results for recognition of Russian vowel phonemes

Test set for each of 5 speakers: 1000 vowels: 100 per each of *R=10* class



The best phoneme recognition results

Distance/features	The best obtained α	Processing time, ms		Accuracy, %	
		SV	FD	SV	FD
Euclidean+MFCC	1	0.7±0.02	1.0±0.01	80±1.7	85±1.4
IS+PSD	0.05	3.5±0.05	5.5±0.04	81.5±1.9	86±1.4
SD+PSD	0.01	1.8±0.03	2.2±0.04	77±1.7	82±1.5

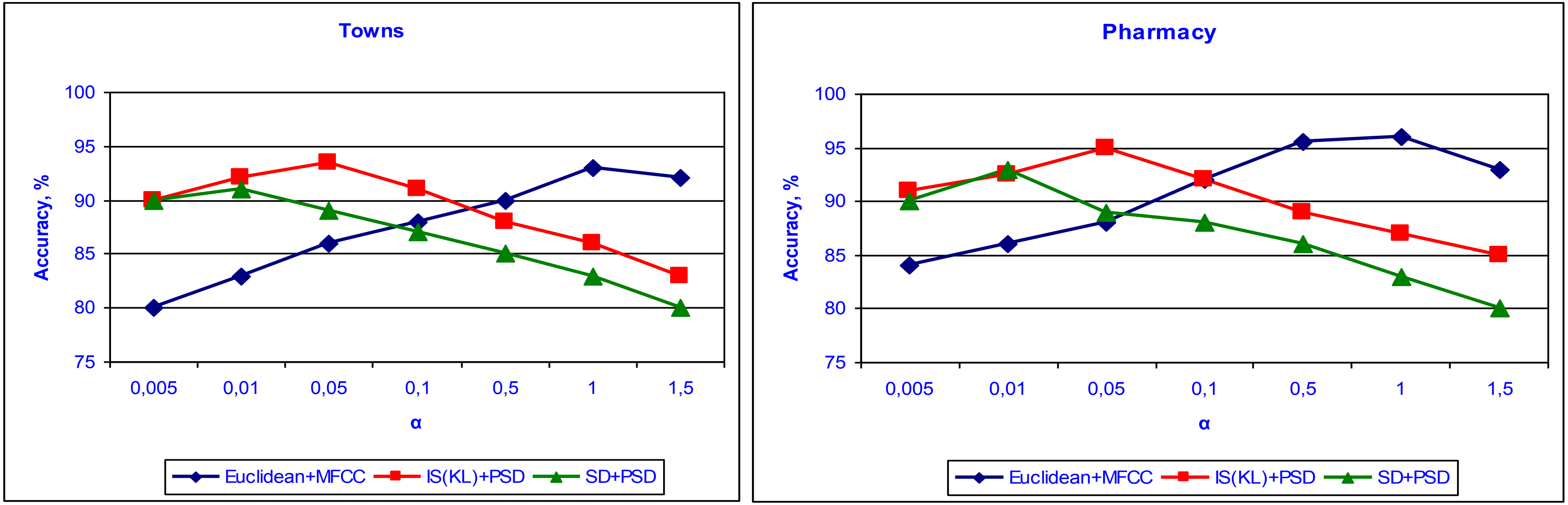
Isolated words recognition task

Test set for each of 5 speakers: 2 words from each vocabulary:

Isolated syllable mode

1. **Pharmacy -** list of 1913 drugs sold in one pharmacy
2. **Towns -** list of 1830 Russian towns with the corresponding region (e.g., “Kstovo (Nizhegorodskaya”))

Dependence of the accuracy on *α*



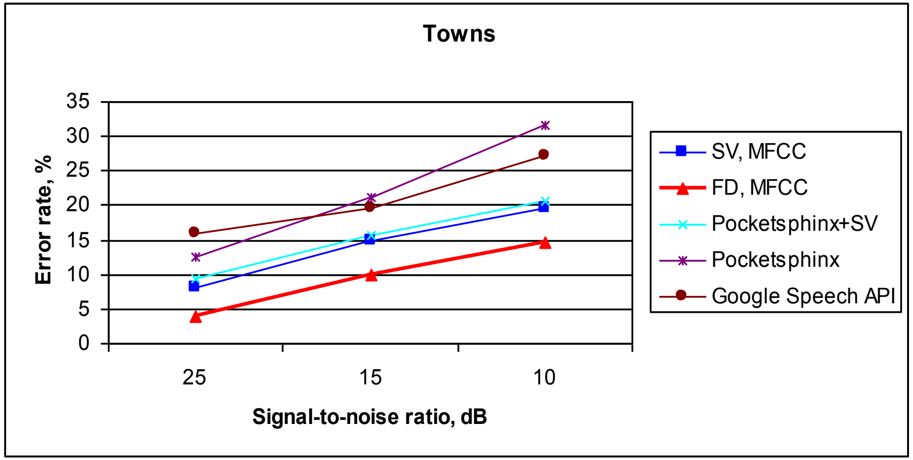
The best accuracy of words recognition is achieved with the same values of *α* as for the phoneme recognition task

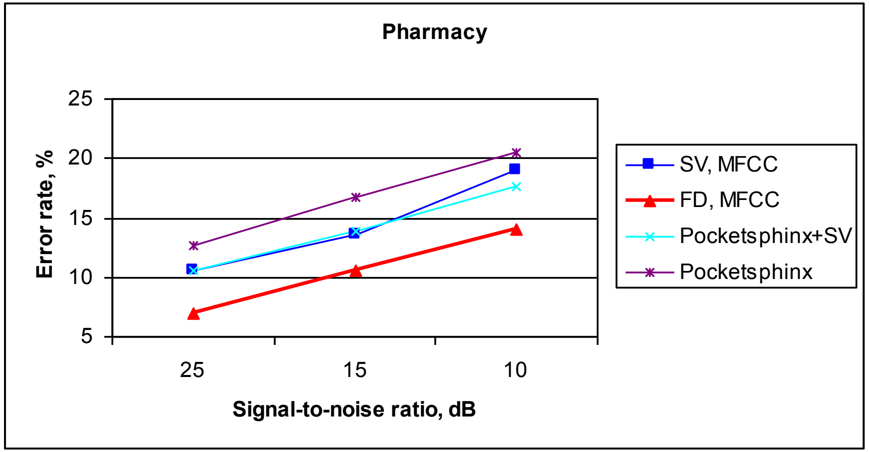
Isolated words recognition task. Comparison with state-of-the-art

Isolated words recognition accuracy, %

Distance/features	Cities		Pharmacy	
	SV	FD	SV	FD
Euclidean+MFCC	92±3.4	**96±2.9**	89.5±2.2	**93±2.0**
IS+PSD	91.5±3.2	**95±3.0**	90±2.0	**93.5±1.9**
SD+PSD	88.5±2.7	**93±2.4**	87±2.9	**91±2.8**
CMU Pocketsphinx (GMM/HMM)	90.5±2.3	-	89.4±3.0	-

Dependence of error rate on the additive noise level





And summarizing our results we have the following conclusions

Fuzzy Decoding method has a list of advantages

1. The usage of the FD method yields to the increase of the recognition accuracy in comparison with conventional voting algorithm
2. The FD method may be successfully applied not only with the Kullback-Leibler discrimination, but with various measures of similarity. For instance, the best recognition accuracy is achieved with state-of-the-art MFCC features comparison in Euclidean space
3. The experiment with Russian speech recognition showed the stability of the smoothing parameter's choice to a type of distance and object features

And disadvantages

1. The computing efficiency of the FD is obviously lower than for the SV technique due to calculation of the posterior probabilities. However, the phoneme recognition time is still reasonable even for real-time applications
2. It is necessary to choice the distance smoothing parameter α properly

Further reading

1. Savchenko A.V. et al. Towards the creation of reliable voice control system based on a fuzzy approach, *Pattern Recognition Letters*, 2015
2. Savchenko A.V. et al. // Proc. of Int.Conf. joint rough set symposium (JRS 2014), *LNCS/LNAI*, 2014.
3. Savchenko L.V. et al. // Proc. of Int.Conf. on nonlinear speech processing (NOLISP 2013), *LNCS/LNAI*, 2013.

What we are going to do in the future

Further research direction

1 Application of the FD method to continuous speech recognition

1.1 Fusion of our vowel recognition with speaker-independent systems (e.g., Pocketsphinx)

1.2 Proper choice of speaker's phonetic database. Speaker adaptation

2 Proposed algorithm adaptation for other set of objects' recognition tasks

2.1 Still-to-still video-based face recognition

2.2 Audio-visual speech recognition

Thank you for your attention

Any Questions?