Optimization of Gain in Symmetrized Itakura-Saito Discrimination for Pronunciation Learning

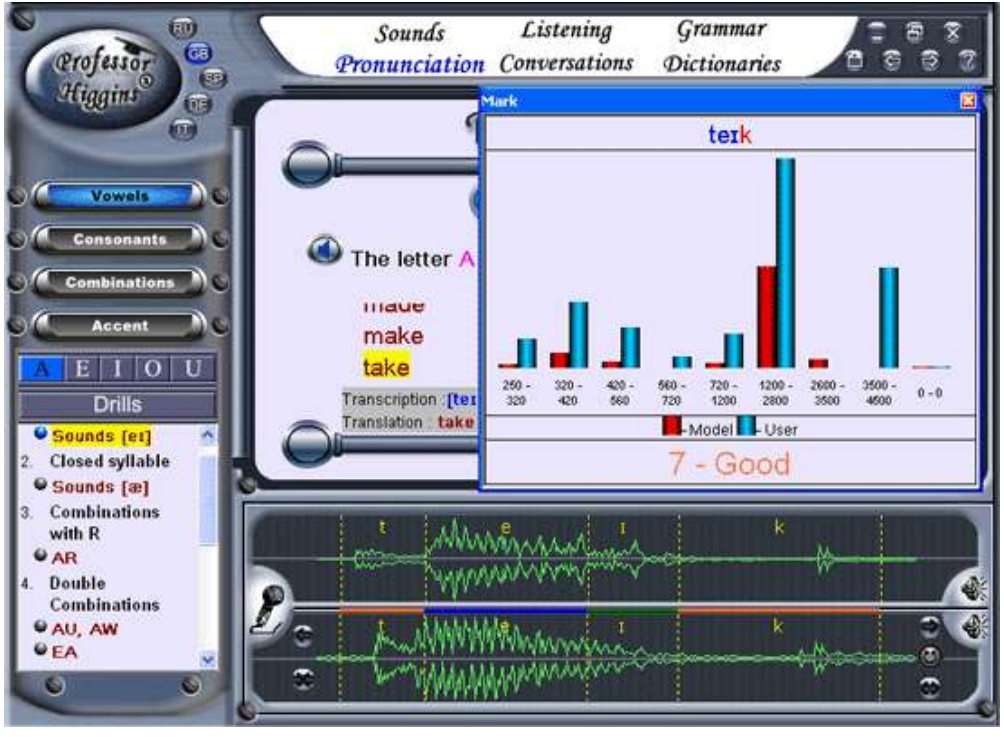
Authors: A.V. Savchenko, V.V. Savchenko, L.V.Savchenko
Presenter: Andrey V. Savchenko
Dr. of Sci., PhD,
Professor in Department of Information Systems and Technology
Senior researcher in LATNA
avsavchenko@hse.ru

Laboratory of Algorithms and Technologies for Network Analysis,
National Research University Higher School of Economics,
Nizhny Novgorod

Computer-aided language learning (CALL) software

What for?

- CALL tool records speech of a user, detects and diagnoses mispronunciations in it, and suggests a way for correcting them.
- One of the most important task is a **pronunciation quality evaluation**.



And now we introduce the agenda of our talk

1. Spectral distortion measures in phoneme recognition
2. Proposed approach
3. Experimental results in phoneme and speech recognition
4. Concluding comments and future works

Learning phoneme pronunciation

- The first task in most CALL systems is to learn correct pronunciation of C ≥ 1 phonemes or short sounds corresponded to letters. The phonetic database of *R* reference phoneme signals $\{\mathbf{x}_r\}$ should be given, and the corresponding label $c(r)$ is known.
- A user learns to produce every sound to be as close as possible to one of ideal signals

$$\min_{r\in\{1,\ldots,R|c(r)=c\}} \rho(\mathbf{x},\mathbf{x}_r) < \rho_c$$

- **Assumption**: speech signals for every sound can be represented as stationary autoregressive (AR) ergodic Gaussian processes with zero mean
- The discrimination $\rho(\mathbf{x},\mathbf{x}_r)$ is typically computed using power spectral densities (PSD).

$$\hat{G}_{\mathbf{x}}(f) = \frac{\hat{\sigma}_{\mathbf{x}}^2}{2F}\left|1+\sum_{m=1}^{p} a_{\mathbf{x}}(m)e^{-i\pi mf/F}\right|^{-2},$$

$$\hat{G}_{r}(f) = \frac{\hat{\sigma}_{r}^2}{2F}\left|1+\sum_{m=1}^{p} a_{r}(m)e^{-i\pi mf/F}\right|^{-2},$$

- F – sample rate
- f=1,2,…,F – discrete frequency
- i – imagery unit
- p – order of the AR order
- $a_{\mathbf{x}}(m)$ and $a_r(m)$, m=1,2,…p – AR/LPC coefficients
- $\hat{\sigma}_{\mathbf{x}}^2$ and $\hat{\sigma}_r^2$ – gains equal to the variance of generative white noise

Spectral distortions

The maximal likelihood solution for testing hypothesis about covariance matrix of the Gaussian signal is achieved by using the Kullback-Leibler (KL) divergence between the zero-mean Gaussian distributions.

It is computed as the **Itakura-Saito (IS) distance** between PSDs:

$$\rho_{IS}(\hat{G}_{\mathbf{x}}, \hat{G}_r) = \frac{1}{F}\sum_{f=1}^{F}\left(\frac{\hat{G}_{\mathbf{x}}(f)}{\hat{G}_r(f)} - \ln\frac{\hat{G}_{\mathbf{x}}(f)}{\hat{G}_r(f)} - 1\right)$$

strongly correlates with the subjective MOS (mean opinion score) estimate of speech closeness

Disadvantage: gains in the PSD estimates depend on the scale of signals.

Gain normalization (GN)

$$\rho_{gn-IS}(\hat{G}_{\mathbf{x}}, \hat{G}_r) = \rho_{IS}(\hat{G}_{\mathbf{x}}/\hat{\sigma}^2_{\mathbf{x}}, \hat{G}_r/\hat{\sigma}^2_r)$$

Gain optimization (GO): Itakura distance

$$\rho_{I}(\hat{G}_{\mathbf{x}}, \hat{G}_r) = \ln\left(\frac{1}{F}\sum_{f=1}^{F}\frac{\hat{G}_{\mathbf{x}}(f)/\hat{\sigma}^2_{\mathbf{x}}}{\hat{G}_r(f)/\hat{\sigma}^2_r}\right)$$

COSH (symmetrized IS) **distance**

$$\rho_{COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_r) = \frac{2}{F}\sum_{f=1}^{F}\frac{(\hat{G}_{\mathbf{x}}(f) - \hat{G}_r(f))^2}{\hat{G}_{\mathbf{x}}(f)\hat{G}_r(f)}$$

Gain normalization is easy for COSH. How to implement gain optimization?

Gain-optimized COSH distance

Gain optimization

$$\rho_{go-COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_r) = \min_{\lambda>0} \rho_{COSH}(\hat{G}_{\mathbf{x}}, \lambda\hat{G}_r),$$

$$\rho_{COSH}(\hat{G}_{\mathbf{x}}, \lambda\hat{G}_r) = \frac{2}{F}\sum_{f=1}^{F} \frac{(\hat{G}_{\mathbf{x}}(f) - \lambda\hat{G}_r(f))^2}{\hat{G}_{\mathbf{x}}(f)\cdot\lambda\hat{G}_r(f)}.$$

$$\frac{d\rho_{COSH}(\hat{G}_{\mathbf{x}}, \lambda\hat{G}_r)}{d\lambda} = \frac{2}{F}\frac{d}{d\lambda}\sum_{f=1}^{F}\left(\frac{\hat{G}_{\mathbf{x}}(f)}{\lambda\hat{G}_r(f)} + \frac{\lambda\hat{G}_r(f)}{\hat{G}_{\mathbf{x}}(f)} - 2\right) = -\frac{2}{F\lambda^2}\sum_{f=1}^{F}\frac{\hat{G}_{\mathbf{x}}(f)}{\hat{G}_r(f)} + \frac{2}{F}\sum_{f=1}^{F}\frac{\hat{G}_r(f)}{\hat{G}_{\mathbf{x}}(f)} = 0.$$

$$\lambda^* = \sqrt{\frac{\sum_{f=1}^{F}\frac{\hat{G}_{\mathbf{x}}(f)}{\hat{G}_r(f)}}{\sum_{f=1}^{F}\frac{\hat{G}_r(f)}{\hat{G}_{\mathbf{x}}(f)}}}.$$

Proposed gain-optimized COSH distance

$$\rho_{go-COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_r) = \frac{1}{F}\sqrt{\left(\sum_{f=1}^{F}\frac{\hat{G}_{\mathbf{x}}(f)}{\hat{G}_r(f)}\right)\left(\sum_{f=1}^{F}\frac{\hat{G}_r(f)}{\hat{G}_{\mathbf{x}}(f)}\right) - 1}.$$

Advantages

- Non-negativity.
- Symmetry.
- Dependence on the ratio of PSDs only.
- Despite the gain-normalized version, the proposed distortion does not depend on scale: every PSD may be scaled without any affect.
- Due to the equivalence of the IS and the KL divergences, one can use the known asymptotic distribution of the KL divergence between samples from the same chi-squared distribution

As the condition is tested by assuming that the input signal represents the c-th sound (correct null hypothesis), the threshold ρ_c can be set to be proportional to the α-quantile of the chi-squared distribution with p(p + 1)/2 degrees of freedom:

$$\frac{\chi^2_{\alpha,p(p+1)/2}}{4(n(\mathbf{x})-p)}$$

Proposed pronunciation learning algorithm

1: **for** $c \in \{1, ..., C\}$ **do** ▷ Learn isolated sounds
2: $N_{reliable} := 0, N_{attempts} = 0, X_c := \{\}$
3: **while** $N_{reliable} < N_{\min}$ AND $\frac{N_{reliable}}{N_{attempts}} < \delta_{\min}$ AND $\bar{r}(X_c) > \bar{r}_0$ (14) **do**
4: $N_{attempts} := N_{attempts} + 1$
5: Record speech signal $\mathbf{x}$ for the c-th sound
6: Compute PSD $\hat{G}_{\mathbf{x}}$ (2) of signal $\mathbf{x}$
7: **for** $r \in \{1, ..., R | c(r) = c\}$ **do**
8: Compute distance $\rho_{go-COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_r)$ (13)
9: **if** $\rho_{go-COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_r) < \rho_c$ **then**
10: $N_{reliable} := N_{reliable} + 1$
11: Append signal $\mathbf{x}$ to the set X_c
12: Break
13: **end if**
14: **end for**
15: **end while**
16: (Optional) add the best utterance (17) to the dataset of reference sounds $\{\mathbf{x}_r\}$
17: **end for**
18: **repeat** ▷ Quality control: recognize isolated sounds
19: $A := 0$
20: **for** $c \in \{1, ..., C\}$ **do**
21: **for** $n \in \{1, ..., N\}$ **do**
22: Record speech signal $\mathbf{x}$ for the c-th sound
23: Compute PSD $\hat{G}_{\mathbf{x}}$ (2) of the signal $\mathbf{x}$
24: Obtain the nearest neighbor r^* (16)
25: **if** $c(r^*) = c$ **then**
26: $A := A + 1$
27: **end if**
28: **end for**
29: **end for**
30: Compute accuracy $A := A/(CN)$
31: **until** $A > A_0$
32: (Optional) Repeat quality control (Steps 18-31) for a sequence of isolated syllables
33: (Optional) Repeat quality control for DNN-based recognition of words

"Radius" for step 3

$$\bar{r}(X_c) = \frac{1}{|X_c|} \sum_{\mathbf{x} \in X_c} \rho_{go-COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_{\mathbf{x}_c^*}), \quad (14)$$

$$\mathbf{x}_c^* = \underset{\mathbf{x}^* \in X_c}{\operatorname{argmin}} \sum_{\mathbf{x} \in X_c} \rho_{go-COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_{\mathbf{x}^*}).$$

The nearest neighbor rule from step 24

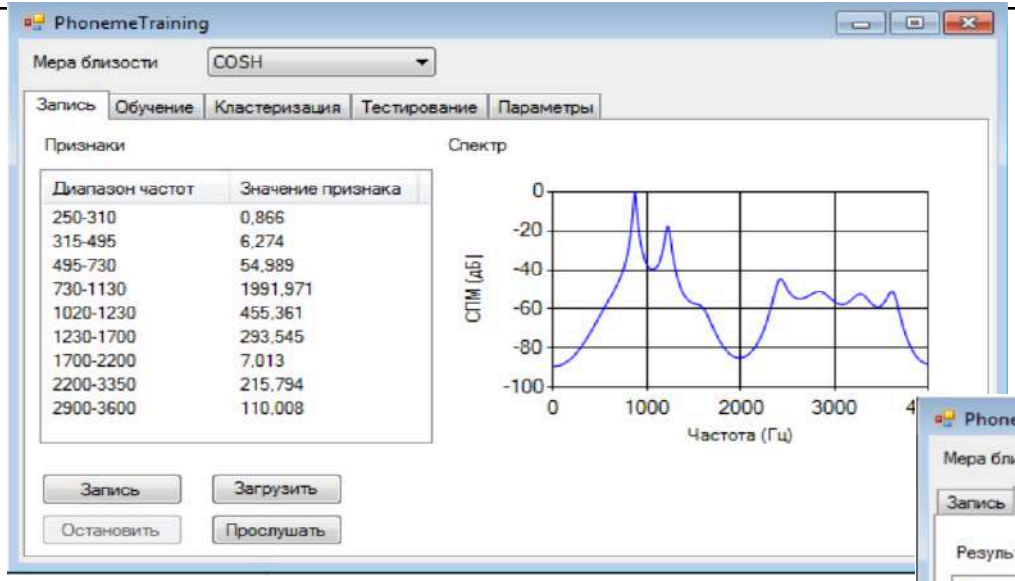
$$r^* = \underset{r \in \{1, ..., R\}}{\operatorname{argmin}} \rho_{go-COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_r). \quad (16)$$

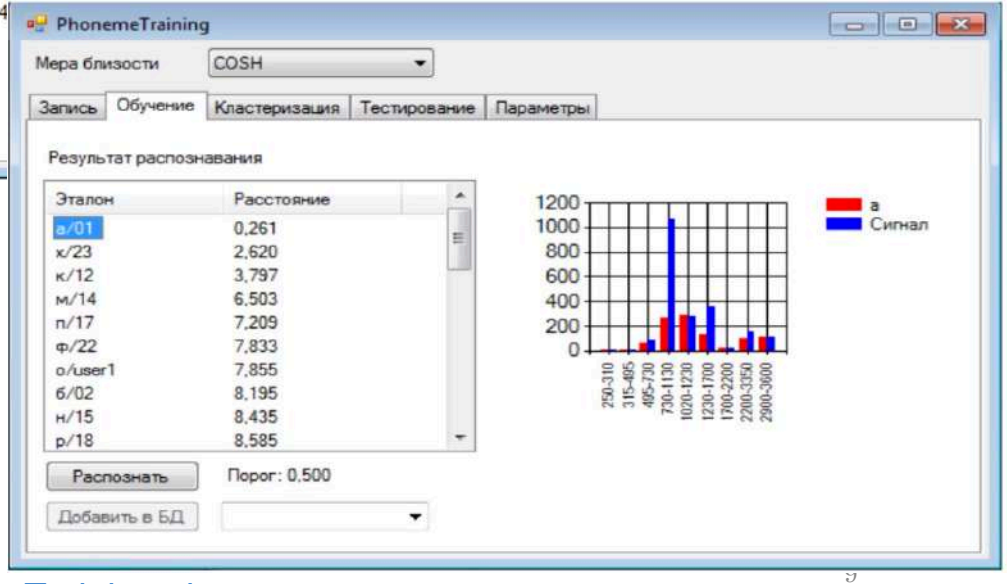
Best-pronounced utterance is added to the dataset of reference sounds at step 16

$$\underset{\mathbf{x} \in X_c}{\operatorname{argmin}} \min_{r \in \{1, ..., R | c(r) = c\}} \rho_{go-COSH}(\hat{G}_{\mathbf{x}}, \hat{G}_r) \quad (17)$$

- PSD is wrapped into the Mel-frequency scale Mel(f) = 1125ln(1 + f/100)
- Weighted sum of PSD samples is computed at regular intervals.
- If speech ranges (f ∈ [200, 3400]) are analyzed and a duration of a regular interval is equal to 55 Mels, such procedure will output only 31 spectral features.

Demo application





https://sites.google.com/site/frompldcreators/PhonemeTraining.zip

Phoneme verification.

- 10 English letters (“a”, “e”, “i”, “j”, “o”, “r”, “u”, “w”, “x”, “y”) pronounced by ideal English native speaker from BBC for CALL software “Professor Higgins”
- 6 Russian vowels pronounced by ideal native speaker for Russian version of “Professor Higgins”
- 5 Russian speakers (3 men and 2 women) produced 1200 isolated vowels (200 for each sound).

Dependence of AUC (area under curve) on SNR (signal-to-noise ratio)

Language	Distance	Signal-to-noise ratio, dB					
		26	20	16	14	12	10
Russian	“Professor Higgins” [5]	88.0	87.1	86.5	86.1	85.3	84.2
	IS (4)	64.1	62.8	61.3	59.3	57.1	55.4
	COSH (5)	79.5	76.6	74.2	71.9	71.4	70.6
	Gain-normalized IS (6)	92.7	92.4	91.1	89.0	86.4	84.0
	Itakura (8)	91.6	90.6	89.5	88.4	87.0	85.4
	Gain-normalized COSH (7)	94.7	94.3	93.2	91.9	90.2	88.7
	Proposed optimized COSH (13)	94.8	94.4	93.4	92.3	90.9	89.4
English	“Professor Higgins” [5]	76.7	77.1	76.8	77.3	75.3	72.8
	IS (4)	73.2	73.1	70.9	68.2	65.2	64.6
	COSH (5)	74.7	76.7	73.9	68.9	67.5	67.9
	Gain-normalized IS (6)	75.6	79.6	77.8	77.8	76.5	76.0
	Itakura (8)	78.2	79.1	77.8	78.1	76.9	76.4
	Gain-normalized COSH (7)	80.8	79.3	79.2	78.9	77.4	76.8
	Proposed optimized COSH (13)	80.7	79.4	79.3	79.0	77.8	77.4

Phoneme recognition

Dependence of accuracy on SNR

Language	Distance	Signal-to-noise ratio, dB					
		26	20	16	14	12	10
Russian	"Professor Higgins" [5]	72.2	71.6	70.2	67.2	64.8	62.3
	IS (4)	32.1	26.4	19.9	17.5	17.4	17.1
	COSH (5)	32.8	33.0	33.0	32.4	31.0	30.0
	Gain-normalized IS (6)	85.7	81.7	75.7	66.9	59.3	55.5
	Itakura (8)	80.4	79.1	77.9	75.6	69.5	62.0
	Gain-normalized COSH (7)	86.3	84.3	79.4	74.3	65.2	58.3
	Proposed optimized COSH (13)	87.0	85.4	81.7	77.1	70.6	62.9
English	"Professor Higgins" [5]	47.5	45.0	47.5	50.0	47.5	35.0
	IS (4)	42.5	42.5	40.0	37.5	30.0	25.0
	COSH (5)	40.0	42.5	42.5	40.0	42.5	35.0
	Gain-normalized IS (6)	60.0	55.0	55.0	52.5	47.5	37.5
	Itakura (8)	55.0	55.0	52.5	50.5	40.0	37.5
	Gain-normalized COSH (7)	62.5	55.5	55.5	45.0	45.0	37.5
	Proposed optimized COSH (13)	62.5	57.5	57.5	52.5	52.5	45.0

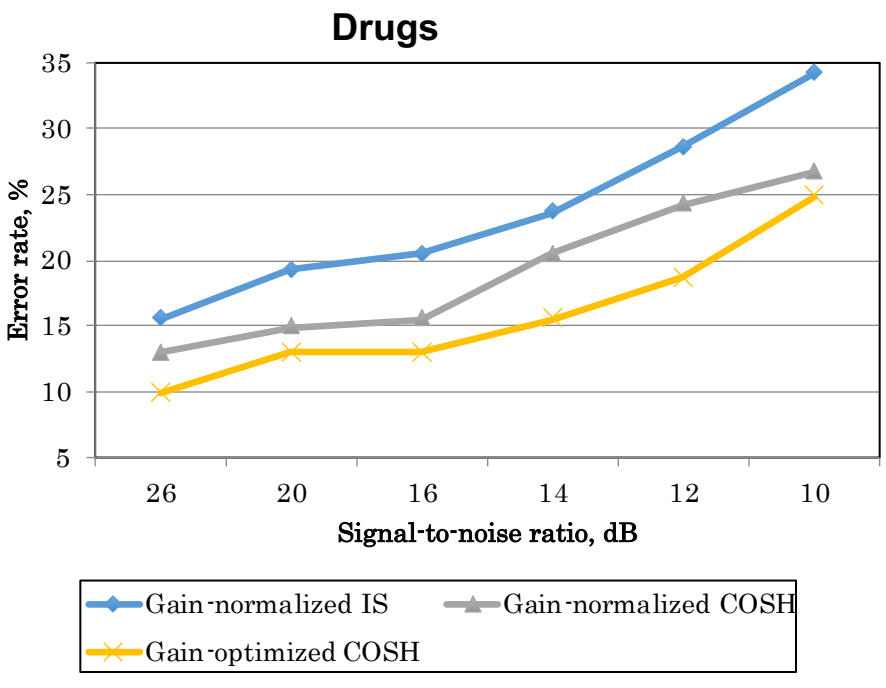
Isolated syllables recognition

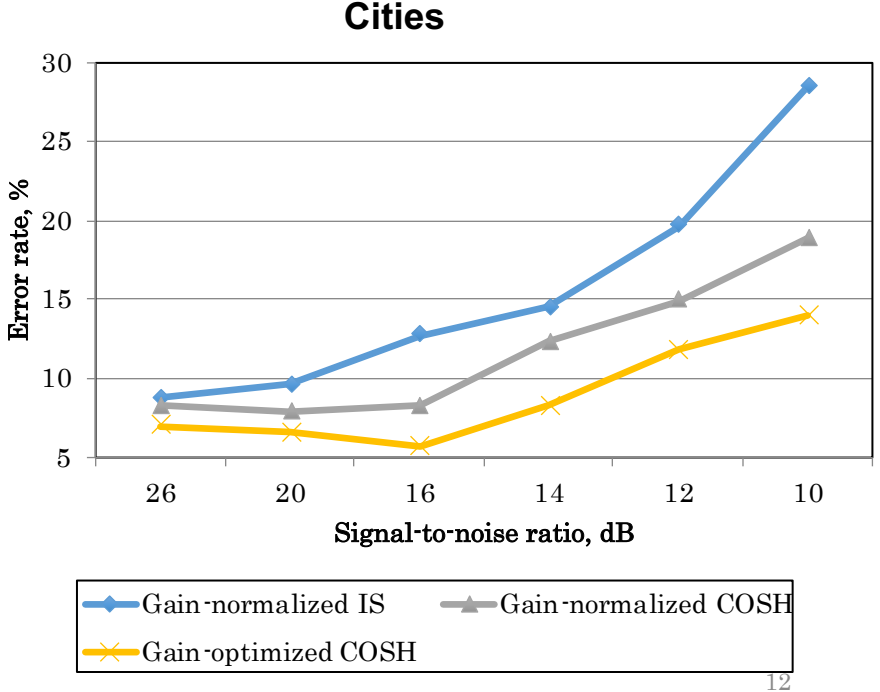
Vocabularies:

Isolated syllable mode

Speaker-dependent mode

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





CNN-based isolated word recognition

- Recognition of the following English words: “down”, “go”, “left”, “no”, “off”, “on”, “right”, “stop”, “up”, “yes”.

- Existing convolutional neural networks (CNN) pre-trained on Google Speech Commands dataset:
  1. “conv” model based on CNN-trad-fpool3;
  2. “low latency conv” model based on CNN-one-fstride4;
  3. “low latency svdf” model with rank-constrained compression;
  4. “tiny conv” model with one fully connected layer.

Average accuracy, %

	conv	low_latency_conv	low_latency_svdf	tiny_conv
Pre-trained	72	27	46	20
Fine-tuned (all words)	91	92	55	50
Fine-tuned (best words)	94	96	60	65

And summarizing our results we have the following conclusions

Proposed approach for CALL systems has several key features

1. Novel spectral distortion based on optimization of the symmetrized Itakura-Saito, i.e., COSH, divergence. It achieves high AUC for pronunciation learning and accuracy for quality control even in the presence of noise in the input utterance.
2. Stability of correct pronunciation is controlled by computing average distances between several recent attempts to pronounce each sound
3. Steps for recognition of sounds and words are introduced to control speech intelligibility
4. We let a user to add the best pronounced sound to the dataset of reference sounds in order to memorize the best attempts and use them in subsequent stages.

What we are going to do in the future

Further research direction

1. Use more complex DNN and LSTM models for speech recognition to control the quality of word's pronunciation.
2. Study speaker adaptation techniques in order to fine-tune the contemporary neural models given the best utterances of a student.
3. Use the proposed algorithm on extra datasets of other languages to analyze its performance and robustness thoroughly.

Thank you for your attention

Any Questions?