PCA-based Feature Extraction for Phonotactic Language Recognition

Tomáš Mikolov, Oldřich Plchot, Ondřej Glembek, Pavel Matějka, Lukáš Burget and Jan “Honza” Černocký

Brno University of Technology
Speech@FIT

31. 6. 2010

Overview

- Introduction
- Motivation
- The Idea
- Results
- Conclusion

Introduction

- Language recognition based on phonotactic models of languages is one of the major approaches
  - N-gram models
    - Likelihoods of sample utterances given language specific models are compared
  - SVM based models
    - Discriminative classifiers are used as models of languages (one against all strategy)
    - Usually, linear kernel and soft margin are used

Motivation of this work

- Exponential growth of number of features: $|V|^n$
  - $|V|$: size of ”vocabulary” (phoneme set), n: n-gram order
  - Usually, $|V| = 30 - 50, n = 3 - 4$
  - In the worst case, we have to deal with more than one million of possible features!

Current approaches

- Discarding of features is based on
  - Frequency
    - Many phoneme combinations never occur in real data
    - N-grams occurring less than some threshold
  - Discriminative information
    - Some n-grams are more important for good classification than others
    - Certain rare n-grams can be actually very useful for classification

Feature vectors for SVMs - the idea

3-gram feature	expected count
a a a	0.1
a a b	0
a a c	0
..	
i n g	0.3
i n k	0.3
y n k	0.3
y n g	0.3
..	

- Many n-grams have high probability of co-occurrence in samples
- Do we really need all these combinations?

Feature extraction using PCA

- Projection of feature space to lower dimensional space seems like natural way to fight curse of dimensionality
- Similar idea already works in n-gram language modeling (neural net LMs)

- Why PCA?
  - Fast, simple, general & well-known
  - Almost no parameter tuning
  - Data driven

Task specification

- NIST LRE 2009 task
- 23 languages, closed set
- 30s, 10s, 3s duration

Feature selection

Performance of 4-gram system with **frequency** based feature selection:

feature size	DEV Cavg 30s
5 000	4.0
10 000	3.5
20 000	3.0
40 000	2.8
80 000	2.7

- $|V| = 33, n = 4, |V|^n = 1\,185\,921$
- Conclusion: for maximum accuracy it is good to have as many features as possible

Feature extraction using PCA

Performance of 3-gram system with PCA based feature extraction:

Features	DEV Cavg 30s	Speedup
$\rightarrow 100$	2.83	1080
$\rightarrow 500$	2.43	199
$\rightarrow 1\,000$	2.38	68
$\rightarrow 2\,000$	2.32	20
$\rightarrow 4\,000$	2.28	3.05
$35\,937$ (full)	2.33	1.0

- Almost no loss of performance when reducing feature space to just 500 dimensions
- It would be computationally infeasible to train full 4-gram system without feature extraction step
- Reported speedups are when SVM models are trained by LibSVM; SVMTorch gives very similar results, LIBLINEAR is several times faster, but with worse accuracy

Results with multiple systems

System	Reduction	EVAL Cavg 30s
HU 3gram	$35\,937 \rightarrow 500$	4.0
HU 3gram	$35\,937 \rightarrow 1\,000$	3.86
HU 3gram	$35\,937 \rightarrow 4\,000$	3.85
HU 4gram	$80\,000 \rightarrow 4\,000$	4.09
EN 3gram	$63\,600 \rightarrow 500$	3.50
EN 3gram	$63\,600 \rightarrow 1\,000$	3.48
EN 4gram	$100\,000 \rightarrow 500$	3.64
RU 3gram	$115\,400 \rightarrow 2\,000$	3.26
RU 4gram	$150\,000 \rightarrow 500$	3.37
RU-ALL 3gram	$115\,400 \rightarrow 1\,000$	3.03

- Results are on NIST LRE 2009 evaluation set
- HUngarian, ENglish and RUssian phoneme recognizers were used to generate features
- Training set size is 10K samples for all systems except RU-ALL (49K samples)

Results with multiple systems

System	Reduction	EVAL Cavg 30s
HU 3gram	$35\,937 \rightarrow 500$	4.0
HU 3gram	$35\,937 \rightarrow 1\,000$	3.86
HU 3gram	$35\,937 \rightarrow 4\,000$	3.85
HU 4gram	$80\,000 \rightarrow 4\,000$	4.09
EN 3gram	$63\,600 \rightarrow 500$	3.50
EN 3gram	$63\,600 \rightarrow 1\,000$	3.48
EN 4gram	$100\,000 \rightarrow 500$	3.64
RU 3gram	$115\,400 \rightarrow 2\,000$	3.26
RU 4gram	$150\,000 \rightarrow 500$	3.37
RU-ALL 3gram	$115\,400 \rightarrow 1\,000$	3.03

- Results are on NIST LRE 2009 evaluation set
- HUngarian, ENglish and RUssian phoneme recognizers were used to generate features
- Training set size is 10K samples for all systems except RU-ALL (49K samples)

Results with multiple systems

System	Reduction	EVAL Cavg 30s
HU 3gram	$35\,937 \rightarrow 500$	4.0
HU 3gram	$35\,937 \rightarrow 1\,000$	3.86
HU 3gram	$35\,937 \rightarrow 4\,000$	3.85
HU 4gram	$80\,000 \rightarrow 4\,000$	4.09
EN 3gram	$63\,600 \rightarrow 500$	3.50
EN 3gram	$63\,600 \rightarrow 1\,000$	3.48
EN 4gram	$100\,000 \rightarrow 500$	3.64
RU 3gram	$115\,400 \rightarrow 2\,000$	3.26
RU 4gram	$150\,000 \rightarrow 500$	3.37
RU-ALL 3gram	$115\,400 \rightarrow 1\,000$	3.03

- Results are on NIST LRE 2009 evaluation set
- HUngarian, ENglish and RUssian phoneme recognizers were used to generate features
- Training set size is 10K samples for all systems except RU-ALL (49K samples)

Results with multiple systems

System	Reduction	EVAL Cavg 30s
HU 3gram	$35\,937 \rightarrow 500$	4.0
HU 3gram	$35\,937 \rightarrow 1\,000$	3.86
HU 3gram	$35\,937 \rightarrow 4\,000$	3.85
HU 4gram	$80\,000 \rightarrow 4\,000$	4.09
EN 3gram	$63\,600 \rightarrow 500$	3.50
EN 3gram	$63\,600 \rightarrow 1\,000$	3.48
EN 4gram	$100\,000 \rightarrow 500$	3.64
RU 3gram	$115\,400 \rightarrow 2\,000$	3.26
RU 4gram	$150\,000 \rightarrow 500$	3.37
RU-ALL 3gram	$115\,400 \rightarrow 1\,000$	3.03

- Results are on NIST LRE 2009 evaluation set
- HUngarian, ENglish and RUssian phoneme recognizers were used to generate features
- Training set size is 10K samples for all systems except RU-ALL (49K samples)

Fusion of systems - final results

Fusion of systems	Cavg 3s	Cavg 10s	Cavg 30s
all 3-gram	15.13	5.01	2.39
all 4-gram	15.85	5.0	2.56
3+4-gram	14.94	4.77	2.34
+ RU3-ALL	14.77	4.65	2.25
+ fixed DEV set	14.13	3.86	1.78

- Results are on NIST LRE 2009 evaluation set

Conclusion

- Feature extraction by PCA provides very high speedups both of training and testing phases, which allows systems to be trained on much more data than usually
- Allows fast tuning of parameters or nonlinear kernels

Future work

- Possibility of even bigger feature space reduction by using nonlinear techniques
- PCA can be estimated on subset of all data to obtain further speedup