Exploring Universal Attribute Characterization of Spoken Languages for Spoken Language Recognition

Author:Sabato Marco Siniscalchi, Jeremy Reed, Torbjorn Svendsen, and Chin-Hui Lee

Processor:陳嘉平

Reporter:吳柏鋒

Outline

- Introduction
- UAR-FrondEnd
- VSM-BackEnd
- Experiment

Introduction

- Spoken language is described with a common set of fundamental units defined “universally”

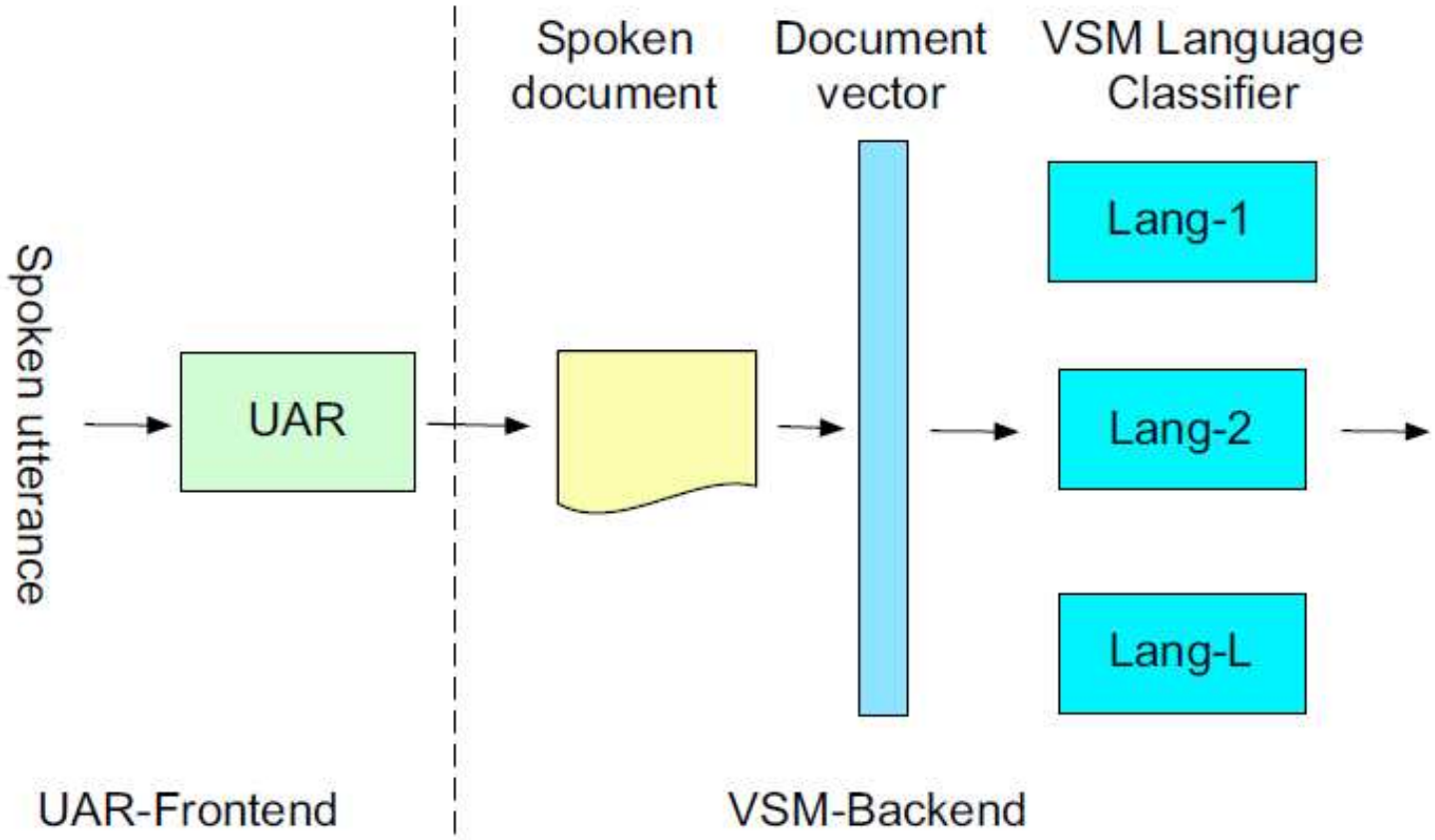
- The advantage of using attribute-based unit is they can be define universally across all language.

Introduction

- Using the vector space modeling (VSM) approaches to language identification (LID) a spoken utterance is first decoded into a sequence of attributes.

- A feature vector consisting of co-occurrence statistics of attribute units is created, and the final LID decision is implemented with a set of vector space language classifiers.

System Overview



UAR-Frondend

- Tokenize all spoken utterances into sequences of speech unit using a universal attribute recognizer (UAR).

- Two phoneme-to-attribute table are created that are phoneme-to-manner and phoneme-to-place.

VSM-Backend

- Manner-based and place-based transcriptions representing speech documents are produced for each speech utterance.

- Each transcription is converted into a vector-based representation by applying LSA.

VSM-Backend

- **LSA (**latent semantic analysis **)**
  is a three step procedure:

  1. term-count vector is created by counting the number of times each term appears in the speech document.

VSM-Backend

2. term-document matrix, $W = \{\omega_{i,j}\}$, consists of weighted count values given by

$$w_{i,j} = \left[1 + \frac{1}{\log N} \sum_{j=1}^{N} \frac{n_{ij}}{n_{i.}} \log \frac{n_{ij}}{n_{i.}}\right] \frac{n_{ij}}{n_{.j}}$$

where n_{ij} is the number of times term i occurs in document j, and $n_{i.}$ is the number of times that term i appears in the N training documents, and $n_{.j}$ is the number of terms in document j.

VSM-Backend

- term-document matrix is quite sparse since many higher order *n*-grams do not appear in training documents.

3. Use singular value decomposition (S.V.D) to reduce the dimensionality and improve the sparsity problem.

Experiment

- The OGI-TS corpus is used to train the articulatory recognizer. This corpus has phonetic transcriptions for six language.

Table 1: *Amount of recorded speech of the OGI-TS corpus in terms of hours per each language.*

Lang.	ENG	GER	HIN	JAP	MAN	SPA	ALL
Train.	1.71	0.97	0.71	0.65	0.43	1.10	5.57
Valid.	0.16	0.10	0.07	0.06	0.03	0.10	0.52
Test	0.42	0.24	0.17	0.15	0.11	0.26	1.35

Experiment

- CallFriend corpus is used for training the back-end language models.

- Test are carried out on the NIST 2003 spoken language evaluation material.

Experiment

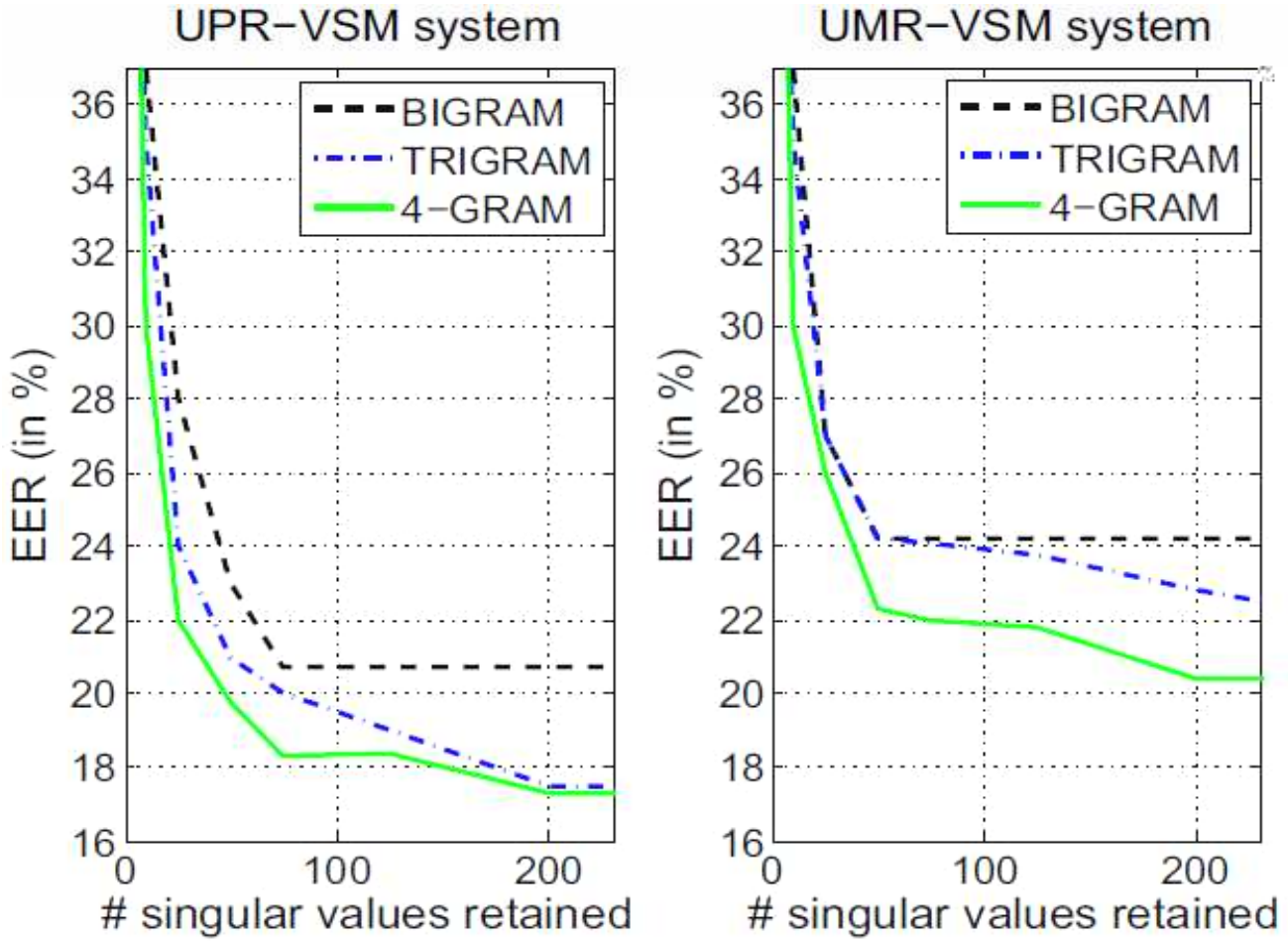
- Language recognition results are reported in terms of equal error rate (EER) , which is the point where the rate of false alarms equals the rate of false rejections.

- Manner-based

  UAR-VSM (UMR-VSM) system
- place-based

  UAR-VSM (UPR-VSM) system

Experiment


UPR−VSM system
UMR−VSM system
BIGRAM
TRIGRAM
4−GRAM
EER (in %)
singular values retained
36
34
32
30
28
26
24
22
20
18
16
0
100
200