INTERSPEECH
2009 BRIGHTON

Exploring Universal Attribute Characterization of Spoken Languages for Spoken Language Recognition

Sabato Marco Siniscalchi[1], *Jeremy Reed*[2], *Torbjørn Svendsen*[1], *and Chin-Hui Lee*[2]

[1]Department of Electronics and Telecommunications, NTNU, Trondheim, Norway
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. USA
{marco77, torbjorn}@iet.ntnu.no, jeremy.reed@gatech.edu, chl@ece.gatech.edu

10.21437/Interspeech.2009-67

Abstract

We propose a novel universal acoustic characterization approach to spoken language identification (LID), in which any spoken language is described with a common set of fundamental units defined "universally." Specifically, manner and place of articulation form this unit inventory and are used to build a set of universal attribute models with data-driven techniques. Using the vector space modeling approaches to LID a spoken utterance is first decoded into a sequence of attributes. Then, 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. Although the present study is just in its preliminary stage, promising results comparable to acoustically rich phone-based LID systems have already been obtained on the NIST 2003 LID task. The results provide clear insight for further performance improvements and encourage a continuing exploration of the proposed framework.

Index Terms: Language recognition, vector space modeling, phonetic features.

1. Introduction

Automatic language identification (LID) is a process of determining the identity of the language spoken in a speech utterance. Broadly speaking, LID approaches can be divided into two main categories: *spectral-based* and *token-based*. The spectral-based approach is purely acoustic and no linguistic information, such as phones or words, is used. Within this context, spoken utterances are represented by sequences of feature vectors, which collectively are used to train a collection of models such as Gaussian mixture models (GMM) (e.g., [1]). In the second category, linguistic properties are exploited in addition to acoustic information. An utterance is first decoded and segmented into a sequence of tokens; e.g., phones. Finally, LID is performed by extracting scores from the resulting token streams. A successful example of this approach is parallel phone recognition followed by language modeling (PPRLM) [2]. It uses several language-dependent phone recognizers to generate phone strings and multiple language-dependent language models to compute phontactic statistics.

As pointed out in [1], the spectral-based paradigm is more efficient and less computationally demanding than the token-based approach, but it does not provide superior performance to the token-based systems on the National Institute of Standards and Technology (NIST) Language Recognition Evaluation (LRE) tasks. The token-based paradigm suffers two main drawbacks: (1) labeled training data is needed to train the recognizers, which is difficult for rarely observed languages or languages without orthography and a well-documented phonetic dictionary; and (2) the decoding phase can be time consuming. To address these issues, several LID systems based on language-independent acoustic phone models have been proposed. For example, [3] builds a collection of 87 phone models from a multilingual telephone corpus, while [4] considers only the phones that best discriminate between languages pairs. Meanwhile, [5] proposes a clustering algorithm in an attempt to deliver a common set of phones for different languages. However, the combined phone list generated from the limited set of initial languages usually does not necessarily cover new and rarely seen languages. A possible solution was reported in [6] where a set of universal acoustic segment models (ASMs) characterizes all spoken languages.

Here we focus on the token-based approach and propose an alternative universal acoustic characterization of spoken languages based on acoustic phonetic features, such as frication, nasalization, etc. In this study, we refer to this set of features as *attributes*. An advantage of using attribute-based units is that they are more fundamental than phonemes, and they can be defined "universally" across all languages [7]. Furthermore, the training material available for several diverse languages can be shared to build a single speech attribute recognizer, which circumvents the problem of needing sufficient labeled data for each language. Meanwhile, using attributes is intrinsically more parsimonious than ASMs. For example, while hundreds of ASMs are needed for a complete characterization of spoken documents [6], the present work uses only 15 attributes. Finally, since the number of these "spoken letters" is small, it is possible to obtain a finer language model resolution such as high order n-grams.

Although this is only a preliminary study on universal attribute characterization of spoken languages for LID, promising results have already been observed. Specifically, the results are comparable to the best performance reported on the 30-second NIST 2003 task when implementing a single phone tokenizer trained on the "stories" part of the OGI Multi-language Telephone Speech (OGI-TS)[1] corpus.

2. System Overview

A bag-of-sounds model can characterize spoken languages similar to the way a bag-of-words model represents documents in the popular latent semantic analysis (LSA) framework [8]. Therefore, LID systems can be realized as in the block diagram in Figure 1 where a front-end processing module tokenizes all spoken utterances into sequences of speech units using a universal attribute recognizer (UAR). This set of acoustic attribute-based symbols represents a collection of shared speech units for

[1] http://cslu.cse.ogi.edu/corpora/corpCurrent.html/

Copyright © 2009 ISCA

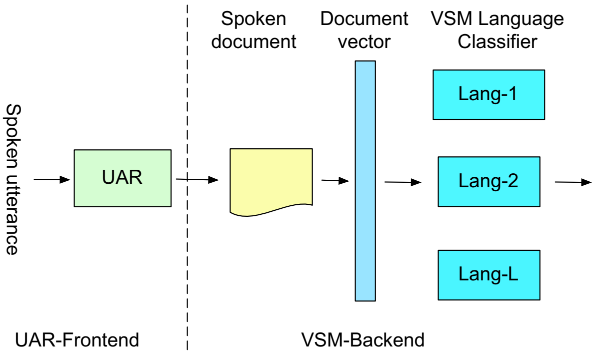

Figure 1: Block diagram of LID system with UAR-frontend and VSM-backend.

all spoken languages, and the sequences of alphabet characters represent the text manifistations of spoken utterances. In addition, it is possible to define acoustic words by grouping multiple acoustic alphabet characters. For example, a single token, two tokens, or up to a sequence of *n>1* alphabet characters can be used as acoustic words. In this study, their co-occurrences are called unigrams, bigrams, and n-grams, respectively. Vector representations of spoken utterances (or documents) are obtained by having each element of the vector characterize the occurrence statistic of an acoustic word (or term). Given a collection of training utterances a term-document matrix is created and text categorization approaches are applied to model each language by considering the training utterances from a corresponding language to form a particular text topic or category. Vector space models (VSM) then categorize unknown utterances into one of a fixed set of spoken languages, i.e., performing the operation of spoken language identification. This VSM-based back-end language classifier is shown in the right-hand side of the block diagram in Figure 1. The UAR-frontend and the VSM-backend are further discussed in the following sections.

2.1. UAR-Frontend

In the present work, manner and place of articulation attributes provide a universal acoustic characterization of all spoken languages. As already mentioned, the main problem of the LID paradigms based on language-independent acoustic phone models is the difficulty in extending the framework to cover new and rarely seen languages. In contrast, speech attributes can be defined "universally" across all languages. Phoneme-to-attribute tables provide a mapping from phoneme transcriptions to attribute transcriptions. Specifically, two phoneme-to-attribute mapping tables were created for all of the six OGI-TS languages; i.e., a phoneme-to-manner mapping table and a phoneme-to-place mapping table. The phoneme-to-manner mapping table has six items: vowel, fricative, nasal, approximant, stop, and silence. The phoneme-to-place attribute consists of ten elements: coronal, dental, glottal, high, labial, low, mid, palatal, silence, and velar.

Once the mapping tables are defined, the OGI-TS phoneme transcripts are converted into two different streams of articulatory attributes, which are used to train, validate, and evaluate both a manner and a place recognizer. These recognizers are built within the hidden Markov model (HMM) framework. Additional details on the design and performance of the recognizers are given in Section 3.2.

2.2. VSM-Backend

At the output of the UAR-Frontend are two sets of attribute-based transcriptions. Specifically, 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 [8]. These document vectors are then used to train vector-based spoken language classifiers (i.e., SVM).

LSA is a three step procedure. First, a term-count vector is created by counting the number of times each term appears in the speech document. A term may consist of a single attribute (i.e., unigram), an ordered pair (i.e., a bigram), an ordered triplet (i.e., trigram), etc. It is here that the manner and place transcriptions are merged by concatentating the manner-based count vector and the place-based count vector for the same utterance. The term-document matrix, $W = \{w_{i,j}\}$, [8] consists of weighted count values given by

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

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. This measure is close to zero if the given term has a uniform distribution throughout the database, but is close to one if the occurrence distribution is skewed to only a few documents.

The term-document matrix has a dimension size of $M \times N$. In general, M equals to the number of unit occurence statistics used, i.e. unigrams, bigrams, trigrams, 4-grams, etc. Therefore $M = p + p^2 + p^3 + p^4$, where p is the number of attributes. For manner and place, this resulted in $M_m = 1554$ and $M_p = 11110$, respectively, for a total of $M = 12664$. Furthermore, the term-document matrix is quite sparse since many higher-order n-grams do not appear in training documents. Therefore, the final step of LSA uses singular value decomposition (SVD) to reduce the dimensionality and improve the sparsity problem. Specifically, the matrix W is decomposed by $W = USV^T$. Retaining only a subset of the largest singular values, converts the word-document space into a lower dimensional "concept" space, where two related documents may have a short distance between them in the reduced space even if they do not have an overlapping term set.

Next, a 1-versus-all multi-class SVM system is trained, such that for an individual target language, a separate SVM is trained with the positive class consisting of the target language and the negative class consisting of all other languages. For LID, one may decide the language identity based on the maximum positive distance from the separating hyperplane [9]. For verification, the method in [6] is employed, where for each target language, a pair of GMMs is determined. The first GMM uses the output SVM distances from the target language utterances to build a target model and the remaining utterances build an anti-target model. The log-likelihood ratio of a given test utterance is compared to a threshold for the final decision.

3. Experiments and Result Analysis

In all the following experiments, all data are conversations recorded over telephone lines, as described in Section 3.1. The articulatory attribute recognition performance is reported in terms of manner error rate (MER) and place error rate (PER) for the manner and place recognizers, respectively. Language recognition results are reported in terms of equal error rate

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

Table 2: *MER on the OGI-TS test sentences.*

System	MLE-SYS	MCE-SYS	ANN-SYS
MER	37.54%	35.59%	27.99%

(EER), which is the point where the rate of false alarms equals the rate of false rejections. All the LID experiments reported in the subsequent sections referred to the 30-second NIST LID 2003 evaluation task [10].

3.1. Corpora

The ”stories” part of the OGI-TS corpus is used to train the articulatory recognizer. This corpus has phonetic transcriptions for six languages: English (ENG), German (GEM), Hindi (HIN), Japanese (JAP), Mandarin (MAN), and Spanish (SPA). For each language, the database is divided into three subsets: training, validation, and test. The overall amount of data in each subset and language is shown in Table1. The training partition of the CallFriend[2] corpus is used for training the back-end language models. It is a collection of unscripted telephone conversation for 12 languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. Each language consists of 20 half-hour telephone conversations for a total of about 10 hours per language. In cases where more than one dialect is available, only one dialect is chosen to train the back-end language models.

Tests are carried out on the NIST 2003 spoken language evaluation material [10]. It is a collection of unscripted telephone conversations of the same 12 languages that are in the CallFriend corpus. All the following LID tests used the 30-second setting, which contains 1280 sessions.

3.2. Attribute Tokenization

Attribute recognizers are designed within the HMM framework. Several alternatives exist to approximate the HMM state probability density functions, and two of those possibilities, namely GMMs, and artificial neural networks (ANNs), are investigated in the following. Specifically, three different manner recognizers are considered. The first is a HMM/GMM system trained with conventional maximum likelihood estimation (MLE). Each of the 6 manner units are modeled with a 3-state HMM, with each state containing a 32-mixture GMM observation density. Spectral analysis is performed using a 22-channel Mel filter bank from 64Hz to 4kHz. Cepstral analysis is then carried out with a Hamming window of 25ms and a frame shift of 10ms, and followed by cepstral mean normalization. For each frame, twelve MFCC features plus the energy coefficient are appended with their first and second time derivatives to yield a 39-dimensional feature-vector. This system is referred to as (MLE-SYS). The second system applies MCE training after building the MLE-SYS seed HMMs and is referred to as MCE-SYS. The HTK toolkit[3] is used to implement these two systems. The third system is a hybrid HMM/ANN system, and it is implemented as in [7], although multi-class ANNs are used rather than binary classifiers. All the ANNs are feed-forward single-layer perceptrons with 500 sigmoidal-based hidden nodes and have a softmax activation function at the output layer. Energy trajectories in Mel-frequency bands, organized in a split-temporal context [11] are used as parametric representations of speech. All of the ANNs are designed using the ICSI QuickNet neural network software package[4], and trained with the classical back-propagation algorithm with cross entropy error function. For all of the three manner recognizers, the training, validation, and evaluation material are used as reported in the last column of Table 1. In Table 2 it lists the MER, in percentage, on the evaluation set. The ANN-SYS system significantly outperforms both the MLE-SYS and the MCE-SYS system. Therefore, this configuration is used to implement both the manner and the place recognizers which tokenize the spoken utterances for the remaining LID experiments reported in Section 3.3. For the sake of completeness, the performance, in terms of attribute error, of the place recognizer is 57.07%.

Table 3: *EER for different UAR-VSM configurations.*

System	PUAR-VSM	PUAR-VSM (R)
EER (in %)	13.5%	11.3%

3.3. NIST 2003 Language Recognition Evaluation

The aim of the first experiment is to find the number of singular values ($|SV|$) to retain in order to achieve a good rank approximation of the term-document matrix, *W*, and reduce the sparsity problem. Figure 2 shows the EER, in percentage, for several values of $|SV|$ for the bigram, trigram, and 4-gram based *W* matrix. In the left panel, results concerning the manner-based UAR-VSM (UMR-VSM) system are shown; i.e., not using the place transcriptions. Similarily, results regarding the place-based UAR-VSM (UPR-VSM) system are shown in the right panel; i.e., not using the manner transcriptions. For the bigram term-document matrix, the EER curves reach a plateau at 50 and 75 singular values retained for the manner and place cases, respectively. In the trigram and 4-gram cases, 200 singular values are needed to achieve a good rank approximation of the *W* matrix. The best EERs are always attained with the 4-gram statistics. Specifically, EERs of 21.5% and 17.5% are obtained with the UMR-VSM and UPR-VSM configurations, respectively.

By emulating the PPRLM idea [2], a parallel attribute recognition followed by a VSM-based language model system was designed (see Section 2.2). Table 3 shows the EERs for parallel UAR-VSM (PUAR-VSM). As expected, the parallel configuration improves LID performance, and a final EER of 13.5% is obtained. To understand how tokenization accuracy affects the performance on the NIST LID 2003 task, the OGI-TS test and training sets (i.e., last column of Table 1) are merged to train the two attribute recognizers. The NIST LID 2003 data is used as the test set. As a confirmation of our intuition, the last column of Table 3 reports a 2.2% absolute error reduction, from 13.5% to 11.3%, with the retrained system (PUAR-VSM (R)).

3.4. Discussion

Language resolution affects the EER as shown in Figure 2. Specifically, as the order of *n*-grams increases, the EER decreases and results in improved LID performance. It is also observed that the EER gain is higher in the manner-related experi-

[2] http://www.ldc.upenn.edu/Catalog/byType.jsp♯speech.telephone

[3] HTK toolkit, http://htk.eng.cam.ac.uk/

[4] ICSI quicknet package, http://www.icsi.berkeley.edu/speech/qn

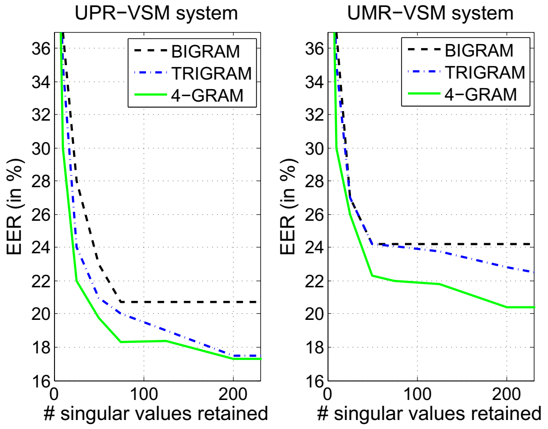

Figure 2: Variation of the percentage of EER on the 30-second NIST 2003 task in terms of retained singular values. In both panels, experimental results for the bigram (*dash line*), trigram (*dashdot line*), and 4-gram (*solid line*) case are shown.

ments than in the place-related ones, but additional experiments with higher order *n*-gram should be performed to draw final conclusions on this issue. A more careful inspection of Figure 2 suggests that the UMR-VSM system suffers from poor acoustic resolution. To validate this conjecture, we used an undirected method: we reduced the number of place classes from 10 to 7 instead of increasing the number of manner phonetic classes. Specifically, the silence and the glottal attributes are folded together into a single class, and the palatal, the dental and the coronal attributes are folded into another single class.

Figure 3 seems to confirm our suspicions. Although the 7-place recognizer has a PER of 45.35% against the 53.07% PER of the 10-place recognizer, the EER drastically increases when a 7-place recognizer is adopted. This further insight suggests that acoustic resolution is more important than recognition accuracy for the LID task. As further confirmation of this assertion, the DET curve 10-place UPR-VSM with bigram statistics is reported in Figure 3. Even with lower language resolution the 10-place solution outperforms the 7-place recognizer with 4-grams. Ways to increase the attribute resolution are currently under study.

The reported results are better appreciated by a qualitative comparison with [11]. In that work, six PRLM systems were trained on the OGI-TS specific language data and tested on the 30-second NIST LID 2003 task. EERs ranging from 11.48% to 15.08% were reported. It was also shown that EER can be drastically reduced when 10 or more hours of specific-language transcribed material is available for training the phone recognizers. In the best case scenario presented in this work, our attribute recognizers are trained with roughly 6 hours of data. We therefore conclude that the proposed UAR-VSM approach to LID is definitely competitive with standard PRLM systems, and we expect to report further improvement in the future by increasing the acoustic and the language resolutions, increasing the training data, and moving from binary SVM classifiers to multi-class classifiers (e.g., ANNs).

4. Summary

This paper proposes a new universal acoustic characterization framework for spoken languages recognition. Based on modeling a shared set of speech attributes, such as manner and place of articulation, a spoken utterance is tokenized into a sequence of universal attribute units so that it can be considered as a spoken document. LSA-based feature extraction and dimension reduction are performed to obtain feature vectors. Vector-based language classifiers are utilized in a similar fashion to designing text categorization systems. By combining manner and place tokenizers we achieve an equal error rate of 11.3% which is comparable with or better than LID systems trained on the same OGI-TS and CallFriend corpora with similar system configurations and complexities. We believe improving attribute transcription accuracy and expanding into multiple attribute tokenizers are two key research directions to enhance attribute-based spoken language recognition system performance.

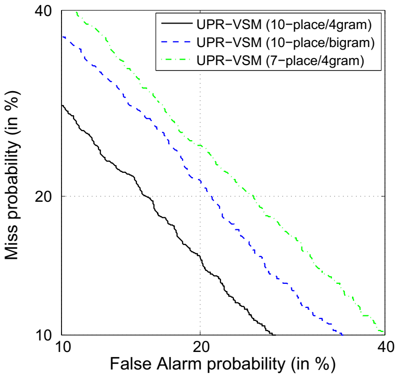

Figure 3: DET plots for several UPR-VSM configurations.

5. References

[1] Torres-Carassquilo, P. A. , Singer, E. Kohler, M. A., Greene, R. J., Reynolds, D. A., and Deller, J. R., Jr., “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” in Proc. of ICSLP, Denver, Colorado, 2002.

[2] Zissman, M. A., “Comparison of four approaches to automatic languages identification of telephone speech,” IEEE Trans. Speech Audio Process., vol. 4 (1), pp. 31-44, Jan. 1996.

[3] Hazen, T. J., “Automatic language identification using a segment-based approach,” M.S. thesis, Mass. Inst. Technol., Cambridge, MA, 1993.

[4] Berkling, K. M., and Barnard, E., “Analysis of phoneme-based features for language identification,” in Proc. of ICASSP, Adelaide, Australia, 1994.

[5] Corredor-Ardoy, C. , Gauvain, J. L., Adda-Decker, M., and Lamel, L., “Language identification with language-independent acoustic models,” in Proc. of Eurospeech, Rhodes, Greece, 1997.

[6] Li, H., Ma, B., and Lee, C.-H., “A Vector space modeling approach to spoken language identification,” IEEE Trans. Audio, Speech, and Lang. Proc., vol. 15 (1), Jan. 2007.

[7] Siniscalchi, S. M., Svendsen, and Lee, C.-H., “Toward a detector-based universal phone recognizer,” in Proc. of ICASSP, Las Vegas, USA, 2008.

[8] Bellegarda, J. R., “Exploiting latent semantic information in statistical language modeling,” Proc. IEEE, vol. 88, no. 8, pp. 1279-1296, Aug. 2000.

[9] Hsu, C.-H. and Lin, C.-J., “A comparison of methods for multi-class support vector machines,” IEEE Trans. on Neural Networks, vol. 13 (2), March 2002.

[10] Martin, A. F., and Przybocki, M. A., “NIST 2003 language recognition evaluation,” in Proc. of Eurospeech, Geneva, Switzerland, 2003.

[11] Matějaka, P., Schwarz, P., Černocký, J., and Chytil, P., “Phonotactic language identification using high quality phoneme recognition,” in Proc. of Interspeech, Lisboa, Portugal, 2005.