



Exploiting Context-Dependency and Acoustic Resolution of Universal Speech Attribute Models in Spoken Language Recognition

Sabato Marco Siniscalchi¹, Jeremy Reed², Torbjørn Svendsen³, and Chin-Hui Lee²

¹Department of Telematics, University of Enna “Kore”, Enna, Italy

²School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

³Department of Electronics and Telecommunications, NTNU, Trondheim, Norway

marco.siniscalchi@unikore.it, jeremy.reed@gatech.edu, torbjorn@iet.ntnu.no, chl@ece.gatech.edu

Abstract

This paper expands a previously proposed universal acoustic characterization approach to spoken language identification (LID) by studying different ways of modeling attributes to improve language recognition. The motivation is to describe any spoken language with a common set of fundamental units. Thus, a spoken utterance is first tokenized into a sequence of universal attributes. Then a vector space modeling approach delivers the final LID decision. Context-dependent attribute models are now used to better capture spectral and temporal characteristics. Also, an approach to expand the set of attributes to increase the acoustic resolution is studied. Our experiments show that the tokenization accuracy positively affects LID results by producing a 2.8% absolute improvement over our previous 30-second NIST 2003 performance. This result also compares favorably with the best results on the same task known by the authors when the tokenizers are trained on language-dependent OGI-TS data.

Index Terms: language identification, latent semantic analysis.

1. Introduction

LID is the process of identifying the language spoken in a sample of speech by an unknown speaker. Each language has its own unique set of characteristics, referred to as the *acoustic signature* of the language, which makes it different from any other language. This acoustic signature can be discovered using information from multiple sources, such as prosody, phonotactic structure, lexical knowledge, acoustic features, vocabulary, and articulatory features. *Spectral-* and *token-based* approaches are statistical techniques usually adopted to decode the acoustic signature of a language. Spectral-based approaches try to determine the language of a spoken query exploiting only acoustic cues (e.g., [1]). Token-based approaches exploit linguistic properties in addition to acoustical information. For example, the phone recognition followed by language modeling (PLRM) [2] approach uses a set of phone models (tokenizer) to convert each speech utterance into a language-dependent string of units (tokens). Multiple interpolated *n*-gram language-dependent models allow higher-order statistics to guide the decision. The parallel phone recognition followed by language modeling (PPRLM) [2] is a successful extension of PLRM.

The token-based paradigm has constantly reported superior results over the spectral-based methods on the NIST Language Recognition Evaluation (LRE) tasks [1]. Unfortunately, the token-based paradigm also suffers two main drawbacks. First, training a tokenizer requires labeled data, which is difficult for rarely observed languages or languages without orthography and a well-documented phonetic dictionary. Second, the decoding phase is usually computationally intensive. To overcome

these issues, several authors have proposed LID systems based on language-independent (or *universal*) acoustic phone models, e.g., [3, 4, 5]. However, the combined phone list generated from the limited set of initial languages usually does not cover new and rarely seen languages. In [6], the authors address this issue by defining a set of universal acoustic segment models (ASMs) that characterize all spoken languages; however, a good characterization requires hundreds of ASMs. Recently, we have presented a novel vector space modeling (VSM) approach to LID [7] that characterizes languages using tokens based on articulatory features. Within this framework, a complete characterization of spoken documents can be obtained using only 15 tokens (5 manner of articulation tokens, 9 place of articulation tokens, and the silence token) compared to the hundreds of acoustic segments used in the ASM-based approach.

In this paper, we extend our VSM-based LID approach by designing better attribute recognizers (tokenizers). Other authors have already shown that the tokenization accuracy directly affects LID results, but this accuracy was improved by simply increasing the amount of labeled training data. Conversely, the key ideas of the proposed approach are to use context-dependent (CD) attribute models, namely right-context (RC) dependent models, and increase the acoustic resolution of the tokenizers. RC subwords models model spectral and temporal information in diverse context better, which produces a more accurate tokenization. Experimental results show that the attribute tokenization error rate decreased from 27.99% to 24.5% and from 57.07% to 36.45% for manner and place, respectively. The set of manner attributes is expanded from five to nine as a first attempt in increasing the acoustic resolution of the manner tokenizer. A final equal error rate (EER) of 8.5% is attained on the 30-second NIST 2003 task when the universal tokenizer is trained using the OGI Multi-language Telephone Speech (OGI-TS)¹ database. This result represents an absolute error reduction of 2.8% with respect to our previously reported performance [7]. To the best of the authors' knowledge, the reported performance also outperforms the best language-dependent PRLM systems trained on language-specific OGI-TS database and tested on the same task [8].

2. Universal Speech Attribute LID System

In [7], we have demonstrated that the LID task can be effectively addressed with the system shown in Figure 1. This LID system consists of two main blocks: a front-end, shown in the left-side of Figure 1, and a back-end, shown in the right-side of Figure 1. The front-end implements a universal attribute recognizer (UAR) that decodes a spoken utterance into two sequences

¹<http://cslu.cse.ogi.edu/corpora/corpCurrent.html/>

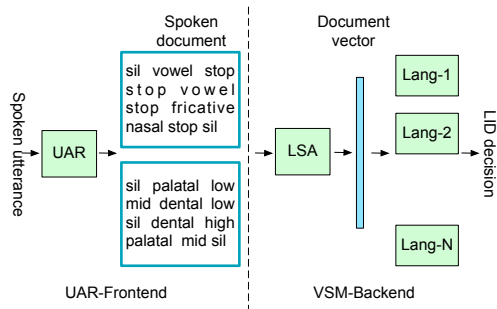


Figure 1: Block diagram of LID system with UAR-frontend and VSM-backend.

Table 1: List of attribute symbols used in this study.

manner	approximant, fricative, nasal, stop, vowel
place	coronal, dental, glottal, high, labial, low, mid, palatal, velar

of universal speech attributes, which have the nice property of being sharable across all spoken languages. The string of attribute symbols (see Figure 1) maps spoken utterances into spoken documents.

The back-end delivers the final LID decision through a VSM approach in two steps. First, a vector representation of the spoken document is obtained using latent semantic analysis (LSA) [9]. The key intuition is that a bag-of-attributes model can characterize spoken languages similar to the way a bag-of-words model represents documents. Specifically, each speech attribute document is represented by a vector of weighted term counts, where a term is an appearance of a single token or an appearance of a sequence of tokens of a certain length, i.e., n -grams. Given a training set of speech attribute documents, the resulting document vectors train a final multi-class classifier used in text document classification by treating each language as a separate topic.

2.1. UAR-Frontend

Universal acoustic characterization of all spoken languages is obtained through manner and place of articulation attributes. These attributes can be defined “universally” across all languages, and phoneme-based transcriptions of labeled audio can be converted into attribute transcriptions using mapping tables. In this work, phoneme-to-manner and phoneme-to-place mapping tables convert phoneme transcripts into streams of articulatory attributes (tokens). The attribute-based transcriptions are used to train, validate, and evaluate UARs. These recognizers are built using hidden Markov model (HMM) technology. The set of attribute units used in this paper are displayed in Table 1. Silence is also taken into account to indicate the absence of articulation activity. The major advantage of the attribute-based characterization over the phoneme-based characterization of spoken utterances is that new and rarely seen languages can be covered without modifying the set of adopted attribute units.

2.2. VSM-Backend

Manner-based and place-based transcriptions are produced for each speech utterance at the output of the UAR-frontend. LSA then converts each transcription into a vector-based representation [9]. Finally, these document vectors train a set of support vector machine (SVM) classifiers.

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 concatenating the manner-based count vector and the place-based count vector for the same utterance. The term-document matrix, $W = \{w_{i,j}\}$, [9] consists of weighted count values given by

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

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 $M \times N$, where M is equal to the number of unit occurrence statistics considered, i.e., unigrams, bigrams, trigrams, 4-grams, etc. In this paper, up to 4-grams are used; 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 to reduce the dimensionality and reduce the sparsity problem. Retaining only a subset of the largest singular values, the word-document space is converted 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 [9].

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. The verification task is accomplished as in [6], 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. Experimental Setup & Results

3.1. Corpora and Performance Metrics

The “stories” part of the OGI-TS corpus is used to train UARs in all experiments. 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 data is divided into three subsets: training, validation, and test. Table 2 shows the overall amount of data in each subset. The VSM-backend is trained on the training part of CallFriend² corpus, which 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 backend. Tests are carried out on the NIST 2003 spoken language evaluation material [10].

²<http://www ldc.upenn.edu/Catalog/byType.jsp#speech.telephone>

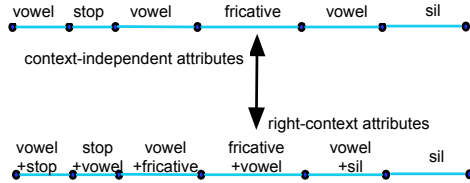


Figure 2: CI-to-RC conversion procedure, an example.

Table 2: 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

This corpus 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. Manner error rate (MER) and place error rate (PER) are the two metrics adopted to evaluate the performance of the universal manner recognizer (UMR) and universal place recognizer (UPR), respectively. The EER, which is the point where the rate of false alarms equals the rate of false rejections, is used to assess LID performance.

3.2. Universal Attribute Modeling

Matějka et al. have already demonstrated that better tokenization accuracies lead to smaller EERs in the PLRM paradigm [8]. A brute-force approach was adopted to achieve this goal, i.e., more labeled training data was used. In this work, the amount of training material is left unchanged with respect to our initial experimental setup [7], and instead, better modeling techniques are exploited to improve attribute tokenization accuracy. Two approaches are explored: (a) attributes are modeled using CD information because CD models have proven to better handle context information, and (b) the acoustic resolution is improved by increasing the number of attribute classes, since it was shown that a finer acoustic resolution results in better LID performance [7].

UARs are implemented within the HMM framework as in [7]. Either Gaussian mixture models (GMMs) or artificial neural networks (ANNs) are used to approximate state probability densities. As a first attempt, RC dependent attribute models are used in place of context-independent (CI) models. There are 36 RC attribute tokens for the manner case, and 100 attribute units for the place case. Therefore, the number of CD models is still quite small, and the data insufficiency is not an issue in the present work. Table 3 lists the MER, in percentage, on the evaluation set for several manner tokenizers. The first system is an MLE-trained RC HMM/GMM system referred to as RC MLE-SYS. Each of the 36 manner units (silence is taken into account as well) is described with a 3-state HMM, with each state containing a 32-mixture GMM observation density. A speech signal is parameterized by a 39-dimensional vector consisting of the first 12 Mel-frequency cepstral coefficients (MFCCs) and the energy term, plus their derivative and accelerations. The second system applies MCE training after building the RC MLE-SYS seed HMMs and is referred to as RC MCE-SYS. The HTK toolkit³ is used to implement these two systems. The third system is a hybrid RC HMM/ANN system (ANN-SYS₅). All ANNs are feed-forward single-layer percep-

³HTK toolkit, <http://htk.eng.cam.ac.uk/>

Table 3: Error rates (in %) on the OGI-TS test sentences. In parenthesis, the number of acoustic attribute units.

	System/Model	CI	RC
Manner	MLE-SYS	37.54 (6)	35.01 (36)
	MCE-SYS	35.59 (6)	33.62 (36)
	ANN-SYS ₅	27.99 (6)	24.5 (36)
	ANN-SYS ₉	–	26.4 (100)
Place	ANN-SYS	57.07 (10)	36.45 (100)

trons with 1000 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 [8] are used as parametric representations of speech. ANNs are trained as in [11]. For ease of comparison with our previous results, the performance of the CI-based systems implemented in [7] are also reported in Table 3. In all cases, the RC-based systems perform better than the CI systems. Moreover, the RC ANN-SYS system significantly outperforms all of the other systems. A absolute 3.5% reduction in the MER with respect to [7] is observed. Specifically, the MER decreases from 27.99% down to 24.5%.

As a first attempt to increase the acoustic resolution of the UMR, the number of manner classes is expanded from 5 to 9 introducing sibilants, affricates, and flaps. The RC HMM/ANN configuration is used, and the MER is reported in the 4th row of Table 3 (ANN-SYS₉). A drop in the tokenization accuracy can be observed when passing from the 5-token to the 9-token configuration, yet this drop is as small as 1.9%.

The RC HMM/ANN configuration is also the solution adopted to implement the UPR. In the last row of Table 3, the performance of this system is reported and compared against the best reported UPR implemented in [7]. By introducing context information in the modeling of the place attribute units, an absolute improvement of 20.62% is achieved in the tokenization accuracy. A final PER of 36.45% is attained. Note that the training phase can start only after having converted CI- into RC-based labels, as shown in Figure 2 for the manner case.

3.3. Language Recognition & Analysis

Figure 3 displays the DET curves using different UARs. In all experiments, the number of singular values retained to achieve a good rank approximation of the term-document matrix, W , and reduce the sparsity problem is 200. Unigram, bigram, trigram, and 4-gram statistics are used in all LID experiments. Figure 3(a) shows the DET curves obtained using only the UMRs. Figure 3(b) refers to DET curves obtained with the UPR only. The DET curves shown in Figure 3(c) are obtained using both the manner and place transcription to compute the statistics of the term-document matrix. DET curves using a CI UAR-frontend are also shown. The DET curves confirm our initial hypothesis; i.e., lower EERs can be obtained by improving the UAR-frontend. An absolute reduction of 4% is achieved when the RC UMR is used, and the EER is reduced to 17.8% from the initial 21.5% obtained using the CI UMR-frontend [7]. Furthermore, Figure 3(a) demonstrates that better LID performance is delivered by improving the UMR acoustic resolution. With this configuration, the EER is further reduced to 16.3%. The EER attained using only the RC UPR is 12.0%, which represents an absolute reduction of 5.5% over the results with the CI UPR-frontend configuration.

A term-document matrix can be generated considering both UMR and UPR transcriptions at the same time. Figure

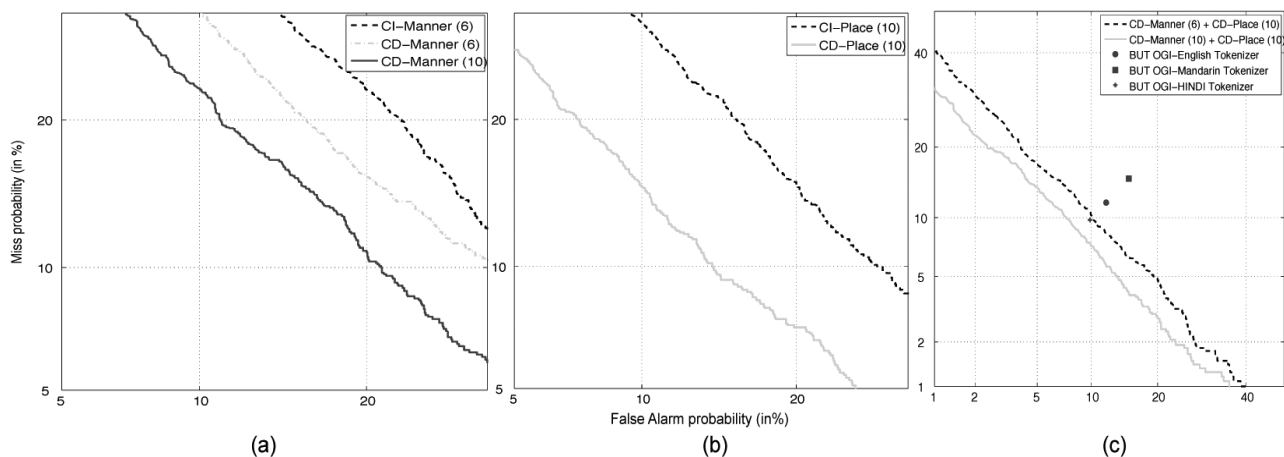


Figure 3: DETs plots for several UAR-VSM configurations.

3(c) shows DET curves obtained with the 6-token RC UMR-frontend (black dashed curve) and the 10-token RC UMR-frontend (green continuous curve). In the first case, an EER of 10% is achieved, and this result is further reduced to 8.5% when the 10-token RC UMR-frontend is adopted.

To better appreciate this latter result, we plot the EERs reported in [8], which were obtained with a language-dependent PRLM approach when the phone recognizer is trained on OGI-TS data. The best and the worst EERs are displayed with cross-shaped and square-shaped dots, respectively. Further, the EER attained with the English-based phone recognizer is reported. Since the green continuous DET curve is below all dots, the proposed approach achieves better results than standard language-dependent PLRM techniques. Moreover, it is worth to notice that the best PRLM result is not delivered with the English-based tokenizer. This quite surprising result demonstrates that more data (see Table 2) does not always imply a better LID performance. Also, PRLM results seem to be strongly tied to the language-dependent available data. Our approach is instead language-independent by design.

The experimental results validate our initial hypothesis, yet a closer inspection of the DET curves reveals an unexpected outcome. Although the 10-token RC UMR outperforms the 10-token RC UPR in terms of recognition accuracy, the UPR-only LID system (EER=12.0%) outperforms the UMR-only LID system (EER=16.0%). There are two main reasons to explain this behavior. Sibilant and affricates are subclasses of the fricative class, and we noticed that they are often confused with stops and fricatives. These errors negatively affect the n-gram statistics of the term-document matrix; e.g., flaps are often deleted. In future work, other classes will be investigated.

4. Summary

An extension of our universal attribute characterization framework for spoken language recognition is proposed. The key idea is to improve tokenization accuracy by incorporating context-information during the attribute modeling process and increasing the acoustic resolution of the tokenizers. EERs of 16.3% and 12.0% were achieved using transcripts generated using the RC UMR-frontend and the RC UPR-frontend, respectively, resulting in approximately a 5% EER reduction over the the LID systems delivered with a CI UAR-frontend (see [7]). By combining manner and place tokenizers, an equal error rate of 8.5% is achieved, which is better than PLRM LID systems trained on the same OGI-TS and tested on the same task. We believe improving attribute transcription accuracy and substituting the

binary language classifiers with a single multi-class language classifier are two key research directions to enhance attribute-based spoken language recognition system performance.

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., Reed, J., Svendsen, T., and Lee, C.-H., "Exploring universal attribute characterization of spoken languages for spoken language recognition" in Proc. of Interspeech, Brighton, UK, 2009.
- [8] Matějaka, P., Schwarz, P., Černocký, J., and Chytil, P., "Phonotactic language identification using high quality phoneme recognition," in Proc. of Interspeech, Lisboa, Portugal, 2005.
- [9] Bellegarda, J. R., "Exploiting latent semantic information in statistical language modeling," Proc. IEEE, vol. 88, no. 8, pp. 1279-1296, Aug. 2000.
- [10] Martin, A. F., and Przybocki, M. A., "NIST 2003 language recognition evaluation," in Proc. of Eurospeech, Geneva, Switzerland, 2003.
- [11] Siniscalchi, S. M., Svendsen, T., and Lee, C.-H., "Toward a detector-based universal phone recognizer," in Proc. of ICASSP, Las Vegas, USA, 2008.