RC25336 (WAT1212-032) December 4, 2012
Computer Science

IBM Research Report

Frame-Based Phonotactic Language Identification

Kyu J. Han, Jason Pelecanos
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 208
Yorktown Heights, NY 10598
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

FRAME-BASED PHONOTACTIC LANGUAGE IDENTIFICATION

Kyu J. Han, Jason Pelecanos

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{kjhan,jwpeleca}@us.ibm.com

ABSTRACT

This paper describes a frame-based phonotactic Language Identification (LID) system, which was used for the LID evaluation of the Robust Automatic Transcription of Speech (RATS) program by the Defense Advanced Research Projects Agency (DARPA). The proposed approach utilizes features derived from frame-level phone log-likelihoods from a phone recognizer. It is an attempt to capture not only phone sequence information but also short-term timing information for phone N-gram events, which is lacking in conventional phonotactic LID systems that simply count phone N-gram events. Based on this new method, we achieved 26% relative improvement in terms of C_{avg} for the RATS LID evaluation data compared to phone N-gram counts modeling. We also observed that it had a significant impact on score combination with our best acoustic system based on Mel-Frequency Cepstral Coefficients (MFCCs).

Index Terms— DARPA RATS, language identification, phonotactic, phone event modeling with timing information

1. INTRODUCTION

Language identification (LID) is the task of identifying which language is spoken for a given recording. Automatic approaches to this problem have been approached from two main directions [1, 2]. One is to utilize frame-based short-term acoustic features such as MFCCs or Shifted Delta Cepstra (SDC) [3] to represent low-level language-specific information in speech. The other is to use phonotactic features derived from one or more phone recognizers to obtain phone N-gram events for the purpose of modeling high-level language-dependent attributes from speech [1, 2, 4]. These two approaches are somewhat complementary and have provided comparable results for the past Language Recognition Evaluations (LREs) hosted by the National Institute of Standards and Technology (NIST).

This work was supported in part by Contract No. D11PC20192 DOI/ NBC under the RATS program. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Recently DARPA launched a program named RATS which studies speech data transmitted through HF radio channels. The RATS program focuses on four tasks; Speech Activity Detection (SAD), Language Identification (LID), Keyword Spotting (KWS), and Speaker Recognition (SR). To evaluate systems under adverse conditions, the Linguistic Data Consortium (LDC) collected the noisy data over eight different channels using a variety of radio transceiver pairs [5].

In this paper we introduce a new approach for phonotactic LID which was a part of our primary system submitted for the RATS LID evaluation. Unlike conventional phonotactic LID systems utilizing phone N-gram event counts, the proposed system exploits frame-level log-likelihoods of phone models during phone recognition and can be used to indirectly capture timing information of phone sequences. To compensate for the potential drawbacks resulting from short-term phone event modeling, it stacks multiple frames of phone log-likelihood ratio features to capture longer-term statistics. In this way phone events can be represented at both frame and segment levels. Adding this indirect timing information to phone event modeling is shown to provide complementary information with both phone N-gram and acoustic feature based systems.

The remainder of this paper is structured as follows. In Section 2, a brief description of conventional phonotactic systems considering phone N-gram events is given, which is followed by presenting our proposed system. In Section 3, we describe the data used for system building and testing, and compare the performance of the two types of phonotactic systems. We also combine the output of the proposed systems with that of our best acoustic system. In Section 4, we wrap up the paper with a summary of our findings.

2. PHONE *N*-GRAM AND FRAME-BASED MODELING

Before we explain our approach to phone event modeling with timing information, we begin by describing a standard phonotactic LID system using phone N-gram event counts.

2.1. *N*-gram count modeling

This method was used in [1, 2, 4] to capture high-level language-specific information from speech regions in audio data. It generally utilizes one or more phone recognizers to convert a stream of acoustic features into a sequence of phone tokens. For each audio recording pre-specified N-gram phone events are counted and the counts are converted to relative frequencies by being normalized by the total phone event counts. The relative frequency of each phone event is then further normalized by background statistics and these features are concatenated to form a supervector. These (per recording) vectors are used for SVM training and testing.

In our paper, we trained an Arabic Levantine open-loop phone recognizer to extract 3- and 4-gram phone event counts. The recognizer was trained in a context-dependent fashion using Perceptual Linear Prediction (PLP) features with Vocal Tract Length Normalization (VTLN) and Feature-based Maximum Likelihood Linear Regression (fMLLR) on the RATS KWS data of approximately 300 hours. We used a dictionary of 50 phone tokens, and 5K context-dependent states for the Hidden Markov Models (HMMs). To obtain the soft counts of the 3- and 4-gram phone events, we used the SRILM toolkit [6], which converted recognition outputs (i.e., phone lattices) into a list of phone events with the corresponding posterior probabilities. The i^{th} element in a phone feature supervector for a given recording X is calculated as $\frac{P(d_i|X)}{\sqrt{P(d_i|B)}}$ [7], where d_i is the i^{th} phone N-gram event of the combined 3- and 4-grams and $P(d_i|\cdot)$ is the relative frequency of d_i for the given recording, i.e., $P(d_i|\cdot) = \frac{\#(d_i|\cdot)}{\sum_{j=1}^{M}\#(d_j|\cdot)}$, where $\#(d_i|\cdot)$ is the smoothed soft count of d_i. B is the background phone event statistics determined from our TRAIN data set (in Section 3.1). For the 4-gram phone events, we selected the 100K most frequent events to make the computational processing more feasible. (There was no pruning for the 3-gram phone events.) The supervector dimension then became $M = 225\text{K} = 50^3$ (3-grams) $+ 100\text{K}$ (pruned 4-grams) . We then applied Principal Component Analysis (PCA) [8] to find a more compact representation. The PCA projection matrix was learned from the aforementioned TRAIN data set to transform the original feature vectors of 225K dimensions into a compactly represented form of 600 dimensions. Using these projected feature vectors, we trained high-order SVM models [9]. For the SVM training, we used the LIBSVM package [10]. As a backend unit, we ran multi-class logistic regression to normalize scores after SVM classification.

2.2. Capturing phone event timing information

Our proposed approach for phone event modeling exploits timing information, which is not captured in conventional methods using phone N-gram counts detailed in Section 2.1.

2.2.1. Frame-based short-term statistics

To indirectly integrate timing information with phone event modeling, we used per frame (10ms each) log-likelihood ratios for the entire phone inventory of the dictionary (50 unigram phones in our case) as an initial feature vector. Consider M (= 50 in our case) phone tokens in a dictionary for a phone recognizer. For each frame f, we generate a feature v_f^m for the m^{th} phone, where $m = 1, 2, ..., M$,

$$v_f^m = \log p(d_m|f) - \log \sum_{i=1, i\neq m}^{M} p(d_i|f). \quad (1)$$

The phone features are collected across phone events and frames to create the set of feature vectors for a recording.

2.2.2. Longer-term phone event modeling

It would be difficult to capture linguistic information with the proposed short-term statistics. In addition, 3- or 4-grams are generally considered better than 1- or 2-grams for phonotactic LID and their statistics cover a longer period in time. In order to capture longer-term phone dynamics, we evaluate two targeted methods.

The first method is to borrow the concept of SDCs [3], which shifts and stacks the original feature vectors with a specified shift (P) and a specified number of stacked frames (K). Considering that on average there are about 10 phones per second in speech, we set P and K such that stacked feature vectors can cover more than 200ms of context. Thus, we selected $P = 4$ and $K = 7$; every 4^{th} frame is stacked until seven frames are stacked in total. (Note that we use the original feature vector elements, not delta information like SDCs, for shift-and-stacking. The original feature vectors in our framework inherently have similar values over adjacent frames so delta features would be of limited benefit.) Once shift-and-stack is performed, the feature vectors have 350 dimensions ($= 50$ phones $\times$ 7 stacked frames). Because of this high dimensionality, a simpler model limitation would need to be imposed.

The second method to consider longer-term phone events is to find a PCA subspace learned from a long-windowed chunk of the frame-based phone feature vectors and to transform such windowed feature vectors to a compact subspace. The PCA transform was determined from a subset of the TRAIN data. The subspace dimension was set to 100 and the window size used for capturing the longer-term context was 1 second.

The resulting feature vectors with longer-term information derived from either the first or second method is then transformed into a fixed dimension supervector by concatenating the normalized MAP adapted mean vectors of a Universal Background Model (UBM) for a recording [11, 12]. After performing supervector MAP adaptation, we applied the same compact representation approach, SVM

training/classification, and backend score normalization as we applied for the conventional phone event modeling with N-gram counts described in Section 2.1.

This proposed method is related to [13] in that it considers timing information (e.g., phone durations) in its modeling of phone events. In [13], duration-specific side information for each phone instance was captured by further categorizing it as a short or long duration event. In contrast, our approach exploits timing information and long-term phone event sequence information by shifting-and-stacking multiple frames.

Very recently a phonotactic system based on phoneme posteriogram counts was introduced [14]. Here the speech was partitioned into segments of phone-like units and for each segment a small vector of the phone unigram probabilities was determined. This information was collected across segment sequences and transformed to represent estimates of N-gram probabilities. We note that while this method does exploit frame based statistics to determine the phone probabilities it does not exploit phone timing information.

3. RESULTS AND DISCUSSION

3.1. Data

The target languages for the RATS LID evaluation are Arabic Levantine, Farsi, Dari, Pushto, and Urdu. LDC distributed the training and development data of the five target languages and ten non-target languages totaling approximately 3,700 hours of recordings (See Table 1). We split the data into three groups for system training, calibration and internal evaluation; TRAIN, COMB, and TEST. The TRAIN data set was used to capture background statistics and train UBMs. This data set was also utilized to find the PCA subspaces. The COMB data set was prepared to calibrate parameters for backend score normalization and score combination. The TEST data set was our internal data used to evaluate the systems. The DEV2 data set is the actual testing data provided by the LDC for the RATS LID evaluation. Table 1 shows the data sets used for our system building and testing in terms of the number of recordings and hours.

3.2. Experimental results

Table 2 shows the performance comparison of a 2-minute duration task[1] for the two systems explained in Section 2. The results are presented in terms of the official metric for the RATS LID evaluation (miss detection rate at 10% false alarm rate, Miss@10%Fa) and the NIST metric for the LREs (C_{avg}). To evaluate the systems, we used both of the TEST and DEV2 data sets as shown. In the 'System' category in the table, 128, 256, and 1,024 indicate the number of Gaussian components in the UBMs[2]. From this table, we observe that the proposed method outperforms the conventional phone event modeling with N-gram counts. The best result was achieved from the PCA-based compact subspace setup for frame-based features with 1,024 Gaussians in the UBM. This system provided a 45% relative improvement compared to the conventional N-gram count approach in terms of Miss@10%Fa for the TEST data set (4.7% $\rightarrow$ 2.6%). In terms of C_{avg} we obtained a 35% relative improvement. A significant improvement was observed in the results for the DEV2 data set as well. The improvements for the DEV2 data set are 42% and 26% (relative) in terms of Miss@10%Fa and C_{avg} respectively.

Table 3 shows the impact of the proposed approach in score combination with our best acoustic system based on MFCCs. The relative improvement of 20% (2.8 $\rightarrow$ 2.3 in Miss@10%Fa) and 10% (6.8 $\rightarrow$ 6.1 in C_{avg}) from the proposed phone event modeling indicates that frame-based phone features perform well by themselves and combine well with acoustic systems.

Table 1. Statistics for the TRAIN, COMB, TEST, and DEV2 data sets for system building and testing.

	No. of Recordings	Hours
TRAIN	87,774	2,926
COMB	9,733	478
TEST	14,328	324
DEV2	1,914	64
Total	113,749	3,792

Table 2. Performance comparison of a 2-minute duration task for the two phonotactic LID systems for the TEST and DEV2 data sets. The numbers in the 'System' column are the number of Gaussian components in the UBMs.

System	TEST		DEV2	
	Miss @10%Fa	C_{avg}	Miss @10%Fa	C_{avg}
N-grams	4.7	8.9	6.9	11.0
Shift-and-stack, 128	3.1	6.7	4.7	8.9
Shift-and-stack, 256	2.9	6.4	4.5	8.5
PCA, 1024	2.6	5.8	4.0	8.1

[1] Training and testing examples are all 2-minute recordings.

[2] The reason that we only used up to 256 Gaussian components in the UBMs in the 'Shift-and-stack' method is because of the high feature dimensionaltiy (= 350) and the corresponding computational resources required. We note that for 256 mixture components the dimensionality of the resulting MAP-adapted supervectors is comparable to the PCA method.

Table 3. DEV2 performance chart including score combination of the two phonotactic systems with our MFCC-based acoustic system for LID. The optimal weights for each combination pair were trained based on the COMB data set.

System	Miss@10%Fa	C_{avg}
1. N-gram	6.9	11.0
2. Proposed (PCA, 1024)	4.0	8.1
1 + 2	3.5	7.3
3. MFCC	2.8	6.8
1 + 3	2.8	6.7
2 + 3	2.3	6.1
1 + 2 + 3	2.3	6.0

4. CONCLUSIONS

In this paper, we proposed a new approach for phone event modeling that aims to exploit timing information for LID. For this, we utilized frame-level log-likelihood ratios of phone models, which are generated from our Arabic Levantine phone recognizer during the decoding process. Using a compact representation of phone likelihood feature vectors for SVM classification, we could achieve 45% relative improvement as compared to the conventional phone N-gram count modeling. In addition, we obtained a significant improvement from the proposed phonotactic LID system in score combination with our MFCC-based acoustic system.

5. ACKNOWLEDGEMENT

We thank Sibel Yaman for providing the high-order SVM utilities and Mohamed Omar for the suggestions.

6. REFERENCES

[1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31–44, Jan. 1996.

[2] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language recognition," *Proc. of Interspeech*, Sept. 1-4, 2003, pp. 1345–1348.

[3] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *Proc. of Interspeech*, Sept. 16-20, 2002, pp. 89–92.

[4] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," *Proc. of Interspeech*, Oct. 4-8, 2004, pp. 1283–1286.

[5] K. Walker and S. Strassel, "The RATS radio traffic collection system," *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, June 25-28, 2012, pp. 291–297.

[6] A. Stolcke, "SRILM - An extensible language modeling toolkit," *Proc. of Interspeech*, Sept. 16-20, 2002, pp. 901–904.

[7] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "High-level speaker verification with support vector machines," *Proc. of ICASSP*, May 17-21, 2004, pp. 73–76.

[8] T. Mikolov, O. Plchot, O. Glembek, P. Matejka, L. Burget, and J. Cernocky, "PCA-based feature extraction for phonotactic language recognition," *Proc. of Odyssey: The Speaker and Language Recognition Workshop*, June 28 - July 1, 2010, pp. 251–255.

[9] S. Yaman, J. Pelecanos, and M. K. Omar, "On the use of nonlinear polynomial kernel SVMs in language recognition," *Proc. of Interspeech*, Sept. 9-13, 2012.

[10] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intel. Systems Tech.*, vol. 2, no. 27, pp. 1–27, 2011.

[11] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[12] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *Proc. of ICASSP*, May 14-19, 2006.

[13] G. Tur, E. Shriberg, A. Stolcke, and S. Kajarekar, "Duration and pronunciation conditioned lexical modeling for speaker verfication," *Proc. of Interspeech*, Aug. 27-31, 2007, pp. 2049–2052.

[14] L. F. D'Haro, O. Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba, and J. Cernocky, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," *Proc. of Interspeech*, Sept. 9-13, 2012.