Proc. Int'l. Conf. Spoken Language Processing, Denver, Sep. 2002 (in press)

Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features[•,♦]

Pedro A. Torres-Carrasquillo[1, 2], Elliot Singer[2], Mary A. Kohler[3], Richard J. Greene[2], Douglas A. Reynolds[2], and J.R. Deller, Jr.[1]

[1]Department of Electrical Engineering, Michigan State University, East Lansing, MI
torresc2@egr.msu.edu, deller@msu.edu
[2]Lincoln Laboratory, Massachusetts Institute of Technology
ptorres@ll.mit.edu, es@ll.mit.edu, greene@ll.mit.edu, dar@ll.mit.edu
[3]Department of Defense USA
m.a.kohler@ieee.org

ABSTRACT

Published results indicate that automatic language identification (LID) systems that rely on multiple-language phone recognition and *n*-gram language modeling produce the best performance in formal LID evaluations. By contrast, Gaussian mixture model (GMM) systems, which measure acoustic characteristics, are far more efficient computationally but have tended to provide inferior levels of performance. This paper describes two GMM-based approaches to language identification that use shifted delta cepstra (SDC) feature vectors to achieve LID performance comparable to that of the best phone-based systems. The approaches include both acoustic scoring and a recently developed GMM tokenization system that is based on a variation of phonetic recognition and language modeling. System performance is evaluated on both the CallFriend and OGI corpora.

1. INTRODUCTION

Automatic language identification (LID) is the process of using a computer system to identify the language of a spoken utterance. Formal evaluations have indicated that the most successful approach to automatic language identification relies on using the phonotactic content of a speech signal to discriminate among a set of languages. Systems based on phonotactic characteristics, such as PPRLM (Parallel Phone Recognition and Language Modeling) [1], typically employ a set of phone recognizers to generate a parallel stream of phone sequences and a bank of n-gram language models to capture the phonotactics. Although phone-based systems provide the best LID performance, their heavy computational demands may preclude their use in low cost, real-time applications. An alternative approach to LID uses Gaussian mixture models (GMMs) to classify languages using the acoustic content of the speech signal. Although GMM systems are quite efficient, they do not provide the superior performance of phone-based LID systems [1]. Recently a variation of the phonotactic approach was proposed [2] in which a Gaussian mixture model, rather than a phone recognizer, was used to tokenize the incoming speech. This approach produced a GMM LID system whose performance was competitive with phone-based approaches but whose operation was much faster.

The present work reports on the performance of GMM-based LID systems that use *shifted-delta-cepstral* (SDC) coefficients as a means of incorporating additional temporal information about the speech into the feature vectors. The use of temporal information spanning a large number of frames is motivated by the success of phonetic approaches that naturally base their tokenization over multiple frames. It will be shown that GMM-based LID systems that use SDC feature vectors perform as well as PPRLM and at a greatly reduced computational cost.

The organization of the remainder of this paper is as follows: Section 2 describes the corpora used for the LID experiments. Section 3 describes LID systems based on GMM acoustic scores and GMM tokenization. Section 4 presents the SDC feature extraction method. Section 5 discusses the LID results obtained for the GMM-SDC LID systems and Section 6 presents conclusions and proposals for future work.

2. CORPORA

The CallFriend corpus [3] is a collection of unscripted conversations for 12 languages recorded over domestic telephone lines. The corpus consists of a training partition used to train the tokenizer and language model components of the systems, a development partition (*devset*) used to train backend classifiers, and an evaluation partition (*evalset*) used to test performance. The 12 languages are: Arabic,

• This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

♦ J.R. Deller was supported in part by the National Science Foundation under Cooperative Agreement No. IIS-9817485. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. P.A. Torres-Carrasquillo was supported in part by The Sloan Foundation and The GE Fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of either The Sloan Foundation or The GE Fund.

English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. Three of the 12 languages (English, Mandarin, and Spanish) contain material for two dialects.

The Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) corpus [4] has been used extensively for the evaluation of LID systems. The training partition of the corpus consists of monologue speech collected from 11 languages, with a total of about 90 minutes per language from different speakers. Each speaker contributed between one and two minutes of speech. The languages are the same languages as for CallFriend, excluding Arabic.

3. GMM-BASED LANGUAGE IDENTIFICATION ALGORITHMS

3.1. GMM LID using Acoustic Scores

The GMM approach to classification has been widely used in a variety of speech recognition applications. The GMM LID system consists of a feature extraction preprocessor, a GMM for each target language, and a backend classifier. Gaussian mixture models trained on speech data from the target language classes produce acoustic, class conditional likelihood scores for each test utterance. GMM LID systems have significant potential advantages over PPRLM since they do not require orthographically or phonetically transcribed speech and are far more computationally efficient. However, performance of GMM-based systems using acoustic scores has tended to be significantly worse than that of the PPRLM system [1].

3.2. GMM Tokenization

The GMM tokenization system [2], shown in FIG. 1, consists of a parallel set of GMM tokenizers, each of which is followed by a bank of tokenizer dependent interpolated (unigram and bigram) language models. Each tokenizer produces a stream of symbols corresponding to the frame-by-frame indices of the highest scoring GMM component. The likelihood of each tokenizer dependent symbol stream is evaluated by the language models, and the language model scores are fed to the Gaussian backend classifier for final processing. The full parallel implementation for the CallFriend evaluation consists of 12 tokenizers, each followed by 12 language models. In the case of the OGI evaluations, only 11 tokenizers are used.

It is apparent from FIG. 1 that GMM acoustic scores are generated by the language tokenizers as a byproduct of GMM tokenization processing. Consequently, these scores may also be appended to the input vector of the backend classifier. This paper will discuss LID results for GMM tokenization with and without appended acoustic scores.

4. SHIFTED-DELTA CEPSTRAL FEATURES

Feature vector extraction for LID systems is typically performed by constructing a feature vector at frame time *t* that consists of cepstra and delta cepstra. However, a

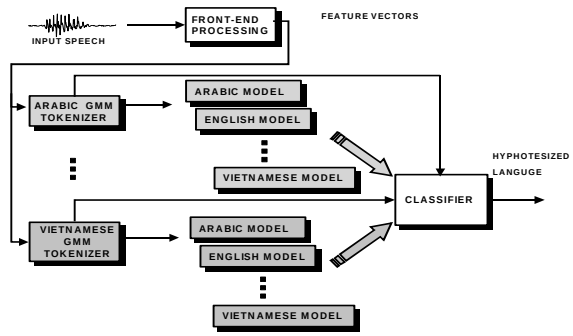

FIGURE 1. Full parallel implementation of the GMM tokenization system.

previous study [5] showed that improved LID performance could be obtained by using shifted delta cepstra (SDC) feature vectors created by stacking delta cepstra computed across multiple speech frames. The computation of the SDC features is illustrated in FIG. 2. The SDC features are specified by a set of 4 parameters, *N*, *d*, *P* and *k*, where *N* is the number of cepstral coefficients computed at each frame, *d* represents the time advance and delay for the delta computation, *k* is the number of blocks whose delta coefficients are concatenated to form the final feature vector, and *P* is the time shift between consecutive blocks. Accordingly, *kN* parameters are used for each SDC feature vector, as compared with *2N* for conventional cepstra and delta-cepstra feature vectors. For example, for the case shown in FIG. 2 the final vector at frame time *t* is given by the concatenation of all the $\Delta c(t + iP)$, where

$$\Delta c(t) = c(t + iP + d) - c(t + iP - d)$$

The SDC features are incorporated into the GMM tokenization system by replacing the conventional cepstral features with SDC feature vectors. The rationale for the use of the SDC features is given in Section 1.

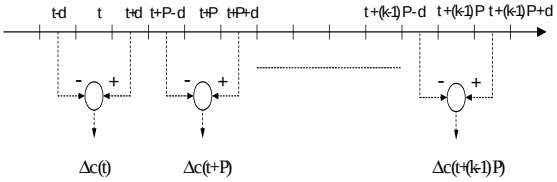

FIGURE 2. Computation of the SDC feature vector at frame *t* for parameters *N-d-P-k*.

5. EXPERIMENTS

This section presents the results of experiments designed to compare the performance of GMM-based LID systems using conventional and SDC feature vectors. GMM LID using acoustic scores was evaluated on the CallFriend corpus while the GMM tokenization system was evaluated on both CallFriend and OGI.

5.1. CallFriend Corpus Evaluation

5.1.1. GMM acoustic scores

LID experiments using GMM acoustic scores were designed to compare the effectiveness of language identification using 1) GMM with conventional cepstra

(GMM-CEP), 2) GMM with shifted delta cepstra (GMM-SDC), and 3) PPRLM. SDC parameterization was set at 10-1-3-3. All systems used Gaussian backends trained from the CallFriend *devset* likelihood scores using diagonal covariance Gaussians with LDA normalization. The dimension of the backend input feature vector was 72 for PPRLM and 12 for GMM. Results (12-language average equal error rate) for the CallFriend *evalset* are shown in Figure 3. The results suggest that replacing conventional cepstra with SDC improves LID performance and that the performance of PPRLM and of high-order (≥512) GMM-SDC are statistically equivalent. Computation time for PPRLM on a Sun SPARCSTATION Ultra-60 is about 3 times real-time (3s of processing for each second of input) and about 0.1 times real-time for GMM512-SDC. The computational load of the GMM512-SDC system is thus about 3% that of PPRLM.

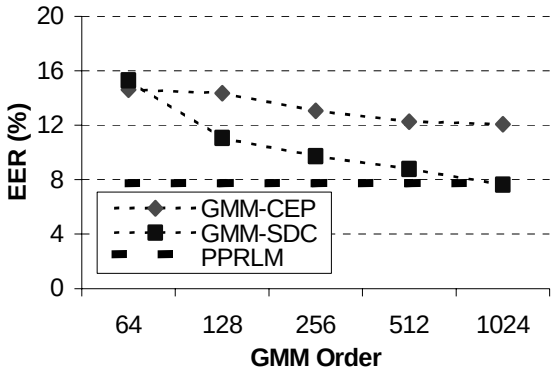

FIGURE 3. LID detection performance (average equal error rate) on CallFriend *evalset* using GMM acoustic scores. 95% confidence intervals are approximately ±1%.

5.1.2. GMM tokenization

The GMM tokenization system was evaluated on the CallFriend corpus under conditions similar to those presented in [2]. A 512-order GMM tokenizer for each language was trained on the CallFriend training partition, which includes 20 conversations per language plus an additional 20 conversations for the dialects of English, Mandarin, and Spanish. The language models were also trained from the CallFriend training partition but without the dialect material. A subset of 1147 test messages from the development partition of the corpus was used to train the Gaussian backend classifier. The full system was tested with the 1492-message CallFriend evaluation partition. The system used 512-order GMM tokenizers, 10-1-3-3 parameterization for the SDC features, and interpolated language models. The SDC parameterization was chosen based on a series of development tests.

The full system was evaluated by running 12 language-dependent tokenizers in parallel. Each of the tokenizers generated an acoustic score and a symbol stream that was scored by 12 language models. Depending on the experiment, a vector of either 144 values (language model scores only) or 156 scores (language model scores plus acoustic scores) was presented to the backend classifier for the final decision.

The plots in FIG. 4 and 5 show the LID performance (12-language average equal error rate) comparison between GMM tokenization systems using conventional and SDC feature vectors as the number of tokenizers was varied. The number of tokenizers was increased by adding one tokenizer at a time from Arabic to Vietnamese, alphabetically, until all 12 were used. In FIG. 4, no acoustic scores were appended to the score vector, and performance seems to be somewhat independent of the tokenizer set and feature vector type. In FIG. 5, the score vector contained both language model and acoustic values and results show a clear improvement with more tokenizers and SDC feature vectors.

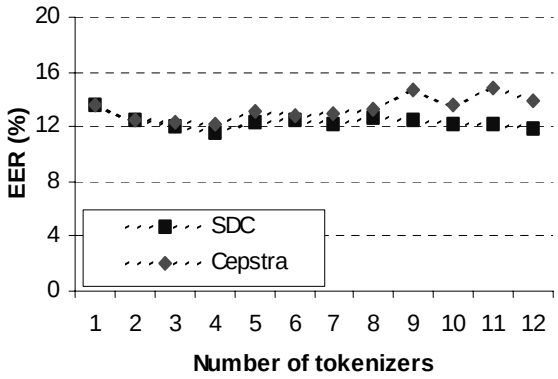

FIGURE 4. CallFriend *evalset* results (average equal error rate) for GMM tokenization (language model scores only) systems using conventional and SDC feature vectors.

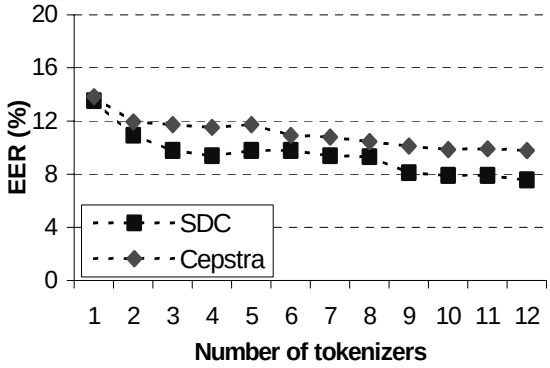

FIGURE 5. CallFriend *evalset* results (average equal error rate) for GMM tokenization (language model and acoustic scores) systems using conventional and SDC feature vectors.

The bar chart of FIG. 6 compares LID performance (12-language average equal error rate) for 4 systems: 1) 512-order GMM-SDC with acoustic scores only ("Acoustic"), 2) 512-order GMM-SDC with language model scores only ("TOK"), 3) 512-order GMM-SDC with language model scores and acoustic scores ("Fusion"), and 4) PPRLM. For this experiment, material from all available CallFriend training messages (including dialects) was used for training tokenizers and language models. The results demonstrate that LID performance comparable to that of PPRLM can be achieved using a 512-order GMM-SDC system.

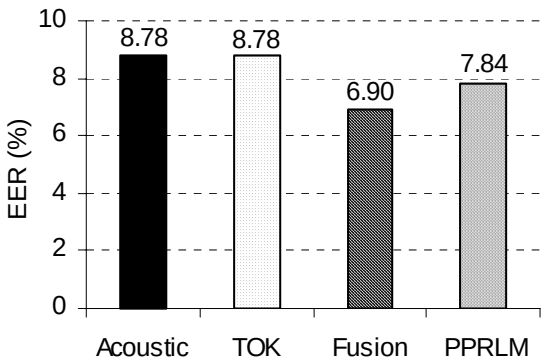

FIGURE 6. CallFriend *evalset* results (average equal error rate) for GMM-SDC and PPRLM. All GMM systems used 512-order mixtures and language model and acoustic scores.

5.2. OGI Corpus Evaluation

There are multiple purposes served by evaluating the system using the OGI corpus. First, other researchers have used the OGI corpus for evaluating their LID systems. Second, the corpus presents new challenges for the GMM tokenization system, the most important of which is the limited amount of data available for training. The OGI corpus contains about 90 minutes of speech per language for training compared to about 10 hours for the training partition of CallFriend. Also, the OGI corpus does not include a *devset* partition and this poses potential problems for training the backend classifier.

Similar to the CallFriend experiments, the training segment of the OGI corpus is used to train both the GMMs and the interpolated language models. The testing partition of the OGI corpus includes two subsets, one containing 45-second messages and the other containing 10-second messages.

The parameters of the system are similar to those described in the section of the CallFriend evaluation with the exception of the GMM order, which is set to 128 mixture components. This different choice of model order is necessary to compensate for the fewer training examples since using a higher order could result in unreliable language models.

5.2.1. Leave-one-out test set evaluation

This experiment uses the leave-one-out technique to evaluate the system performance. The use of the leave-one-out technique allows the system error rate to be estimated while maximizing the use of the available training data. The results for this evaluation, shown as closed-set percent error to allow comparison to other published results, appear in FIG. 7.

The performance obtained by the system on the OGI corpus is significantly worse than that obtained by the PPRLM system [1]. There are at least two possible explanations for this decreased performance. First, it has been shown that the system performance increases with the GMM order, since this provides not only better language model scores, but also better acoustic scores. Second, the availability of more examples for training the backend classifier can also help performance, particularly for the 45-second utterance test.

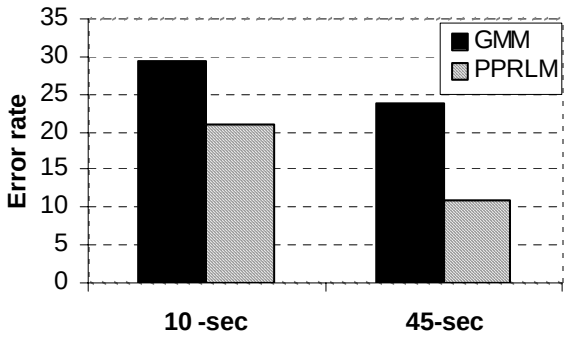

FIGURE 7. OGI test set results (percent error) for the GMM tokenization system and PPRLM.

6. CONCLUSIONS

This paper has presented the latest results obtained in the continuing efforts to develop a more flexible and adaptable system for language ID. The results show significant improvement over the previous system results [2] by the use of shifted-delta cepstra features.

The GMM tokenization system was also evaluated using the OGI corpus to determine its performance and adaptability to new conditions. The evaluation of the system using this corpus has shown higher error rates compared to other systems in the literature [1, 6, 7] but at a lower computational cost and without requiring *a priori* information about the speech such as transcriptions.

The future work includes further analysis of the amount of data needed for training the interpolated language models and backend classifier. Also, future experiments will be conducted to assess the system performance using shorter speech segments and cross-corpus experiments.

7.REFERENCES

[1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, 1996.

[2] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller Jr., "Language identification using Gaussian Mixture Model Tokenization," In ICASSP, Orlando, Fl. , USA, 2002.

[3] CallFriend Corpus, Linguistic Data Consortium, 1996, http://www.ldc.upenn/ldc/about/callfriend.html

[4] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," In ICSLP, Alberta, Canada, 1992.

[5] B. Bielefeld, "Language identification using shifted delta cepstrum," In Fourteenth Annual Speech Research Symposium, 1994.

[6] Y. Yan and E. Barnard, "An Approach to Automatic Language Identification Based on Language-dependent Phone Recognition," In ICASSP '95, Detroit, Michigan, 1995.

[7] J. Navratil and W. Zuhlke, "Phonetic-context mapping in language identification," In EUROSPEECH, Rhodes, Greece, 1997.