interspeech 2008 incorporating SST 08

Politecnico di Torino System for the 2007 NIST Language Recognition Evaluation

Fabio Castaldo [1], *Emanuele Dalmasso* [1], *Pietro Laface* [1], *Daniele Colibro*[2], *Claudio Vair* [2]

[1] Politecnico di Torino, Italy, [2] Loquendo, Torino, Italy

{Fabio.Castaldo, Emanuele.Dalmasso, Pietro.Laface}@polito.it
{Daniele.Colibro, Claudio.Vair}@loquendo.com

Abstract

This paper describes the system submitted by Politecnico di Torino for the 2007 NIST Language Recognition Evaluation. The system, which was among the best participants in this evaluation, is a combination of classifiers based on three acoustic models and on two sets of Parallel Phone tokenizers. It exploits several state-of-the-art techniques that have been successfully applied in recent years both in speaker and in language recognition.

We illustrate the models, the classification techniques and the performance of the system components, and of their combination, in the NIST-07 close-set 30 sec General Language Recognition task. We also highlight the difficulties in setting appropriate decision thresholds whenever the training data of a language are scarce, or the test data are collected through previously unseen channels.

Index Terms: Spoken Language Recognition, LID, Feature compensation, Phone tokenizers

10.21437/Interspeech.2008-138

1. Introduction

In the recent years a substantial reduction of the error rates in spoken language recognition has been obtained [1]. This progress has been achieved by more accurate acoustic and phonetic language models, by introducing inter-speaker and channel compensation techniques [2-5], and by exploiting discriminative training approaches [6-8].

In this paper we present a successful fusion of these techniques in a system that was submitted for the 2007 NIST Language Recognition Evaluation (LRE) [9]. The system is the combination of classifiers based on two sets of Parallel Phone tokenizers exploiting high order multigrams, and on three acoustic models, Phonetic GMMs [2], classical GMMs [10], and the latter in combination with SVM classifiers [5],

In the following, we describe the system components, the training and development databases, and the experimental results obtained in the LRE07 General Language Recognition close-set 30 sec task, highlighting some still open problems such as the quality or the lack of training/test data for a language.

2. Acoustic models

Gaussian Mixture Models used in combination with Maximum A Posteriori (MAP) adaptation represent the core technology of most state of the art text-independent speaker recognition systems [10]. Although it is possible to train reliably a GMM of a language by Maximum Likelihood estimation, due to the large amount of training data usually available, we perform MAP adaptation from a Universal Background Model (UBM) even for language models. The main reason for this choice is that language models deriving from a common UBM are required by our GMM-SVM approach, and by our frame based inter-speaker variation compensation approach [4], which computes its speaker factors using the UBM. Moreover, fast Gaussian selection is performed in training and test using the UBM.

2.1. GMM system

In the experiments described in this paper, the UBM and the language GMMs consist of mixtures of 1024 Gaussians. The observation vector includes 56 parameters: the first 7 Mel frequency cepstral coefficients and their 7-1-3-7 Shifted Delta (SDC) coefficients [11]. Gender dependent UBMs have been trained on the LDC Callfriend corpus, including approximately 450 hours of telephone conversations in 12 languages [12].

To reduce inter-speaker variability within the same language we have shown in [2] that significant performance improvement in LID can be obtained using factor analysis. We estimate an inter-speaker subspace that represents the distortions due to inter-speaker variability, and compensate these distortions in the domain of the features. The details of this approach are given in [2] and [4].

Using compensated features, we trained a gender-dependent model for each of the 12 target languages in the NIST corpora using the training and development sets of the CallFriend corpus. The conversations in this corpus were split into 8172 segments of approximately 150s. The same data sets were used for training all other types of models.

2.2. Phonetic GMM system

The Phonetic GMM (PGMM) system used for LRE07 has the same architecture of the system, described in [2], which was used for the NIST 2006 Speaker Recognition Evaluation. The only difference is that we use the same speaker-compensated Shifted Delta features of the language recognition GMMs.

We decode an utterance, both in enrollment and in recognition, producing phonetic labeled segments. The decoder, described in Section 3, is trained to recognize 11 language independent broad phone classes: silence, liquids, nasals, fricatives, affricates, voiced and unvoiced plosives, diphthongs, front, central, and back vowels. Each broad class is modeled by a single state that has associated a GMM. These models have been trained with 20 hours of speech in 10 different languages,

Emanuele Dalmasso is supported by a Lagrange Project scholarship of the CTR and ISI Foundations, Torino, Italy.

using Macrophone [12] for US English, and the SpeechDat 2 corpora [13] for (Dutch, French, German, Greek, Italian, Portuguese, Spanish, Swedish, and UK English).

The UBM and the language models, obtained adapting the UBM, consist of the union of the phonetic GMMs associated with each phone class. For each state, the maximum number of (diagonal covariance) Gaussians per mixture is 128, and the total number of Gaussians of this system is 1280 (because the silence model is excluded). The UBM for the PGMMs has been trained using the Callfriend corpus, and another 100 hours of telephone speech from Italian, Portuguese, and Swedish SpeechDat 2 corpora.

In enrollment, the labels and the boundaries of the phonetic segments are used for MAP adaptation of the parameters of the gender and class-dependent GMMs. In recognition, the phonetically labeled audio segments are scored against their corresponding GMMs. Thus, the likelihood of a given observation vector is computed by selecting the GMM corresponding to the phone class decoded at that time frame.

Both the gender dependent GMMs and the PGMMs are discriminatively trained by means of Maximum Mutual Information Estimation [7].

The Phonetic UBM is also used as a speech activity detector for all the acoustic systems, by discarding the speech intervals recognized as silence.

2.3. SVMs using GMM supervectors

Gaussian Mixture Models in combination with a Support Vector Machine classifier (GMM-SVM) have been shown to give excellent classification accuracy in speaker recognition [2], and in language identification [6], [8].

For the GMM-SVM approach we doubled the number of mixtures of the models because we estimated gender independent GMMs. Thus, 2048 Gaussian models are obtained by MAP adaptation, with a small relevance factor, from a common UBM trained using the training and development sets of the Callfriend data. A specific GMM is trained for each segment of a language, both in training and in testing. A supervector that maps a segment to a high dimensional space is obtained by appending the adapted mean value of all the Gaussians of a GMM in a single stream, after appropriate rescaling [2]. The normalized supervectors are used as samples for training linear SVM classifiers.

3. Phone models

Since the combination of acoustic and phonetic systems is known to give good performance [7],[14], we exploited the availability of several languages in the Loquendo-ASR recognizer [15] to implement a phonetic system based on the Parallel Phone tokenizer-SVM approach, which was first proposed for speaker recognition [16],[17].

3.1. Parallel Phone tokenizers

The Loquendo-ASR uses a hybrid HMM-ANN model, where each phonetic unit is described in terms of a single or double state left-to-right automaton with self-loops. The models are based on a set of gender independent units, consisting of stationary context independent phones and diphone models. The ANN is a three layer Multilayer Perceptron that estimates the posterior probability of each unit state, given a context window of 7 frames consisting of acoustic feature vectors including 13 RASTA PLP parameters and their first and second derivatives. The ANN has 315 units for the first hidden layer, and 300 for the second hidden layer. Softmax normalization is applied to the output layer, which includes a language dependent number of states (~700 - 1000).

For these experiments, a phone-loop grammar with diphone transition constraints has been used, and the statistics of the n-gram phone occurrences in each segment were collected from the best decoded string only. Each train and test segment has been transcribed by 9 different decoders for the following languages: Catalan, German, French, Italian, Polish, Spanish, Swedish, UK and US English.

3.2. Language recognition SVM

The SVM approach, proposed in [17] for speaker recognition, uses the phone streams of several utterances of a speaker to produce her/his target model. We applied the same technique to language identification [8].

Given a phone sequence produced by a phonetic transcriber, the frequency of the n-grams within the sequence is computed. The frequency of each n-gram is normalized by the square root of its frequency in the whole training set. By appending in a single vector all these normalized n-gram frequencies, we obtain the so called Term Frequency Log-Likelihood Ratio (TFLLR) kernel [17]. A linear SVM model of a target language is trained by using the vectors computed for each segment of the target language as positive examples, and the set of the vectors of all the other language segments as negative examples.

3.2.1. Multigrams

Usually, the TFLLR kernel include all n-grams $n=0,1,...N$ that appear in the sequences of the training set. The total number of the different n-grams (up to order 3) that appear in our training corpus for the 9 language transcribers is shown in line 1 of Table 1. The dimension of the normalized vector is potentially huge, but since we use relatively short segments, the resulting vectors are sparse.

For this evaluation, we used two different TFLLR kernels, the first one based on trigrams, and the second one relying on pruned multigrams. The use of multigrams can provide useful information about the language by capturing regularities of variable length within the sequences [18]. We prune the list of the n-grams appearing in the training set according to a simple criterion. For each phonetic transcriber, we discard all the n-grams appearing in the training set less than a preset percentage of the average occurrence of the unigrams in the training corpus. The threshold has been fixed to 0.05% in these experiments. Line 2 of Table 1 shows the distribution of the number of n-grams resulting by the application of this threshold. It is interesting noting that the total number of trigrams and four-grams is balanced, and that a fair amount of high order n-grams exist that possibly describe words or sub-words.

A single score is produced by a phone model because we include in a single vector the normalized frequencies of the n-grams of all the phonetic transcribers.

4. Training and development data

We trained a gender-dependent model for each of the target languages/dialects in the NIST LRE07 [9] (and a gender-

Table 1. *Total number of different n-grams in the training set for 9 language transcribers.*

N-gram	1	2	3	4	5	6	All
Not-pruned	344	8785	232091	-	-	-	241220
Pruned	344	8525	90451	90979	8995	112	199406

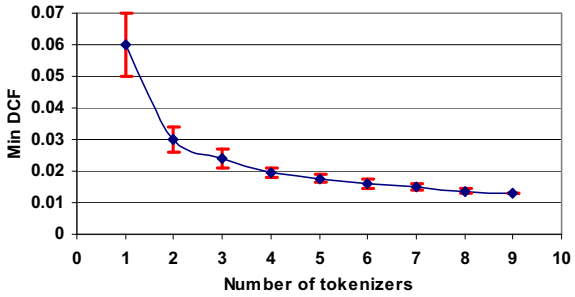

Figure 1. *Range of Min DCFs as a function of the number of fused tokenizers.*

independent model for the GMM-SVM approach) using the following corpora:

- All data of the 12 languages in the Callfriend corpus.
- Half of the NIST LRE07 development corpus.
- Half of the OHSU corpus provided by NIST for LRE05.
- The Russian through switched telephone network [12].
- Since we estimated that the development data provided by NIST for Bengali and Thai were not sufficient, we searched a source of audio file in these languages in the net. We found in [19] such a source, where streams of religious readings excerpts are available. Although these data do not match the conditions of the other corpora because they consist of read, microphone speech collected through an unusual channel, our experiments on the development sets seemed to support the idea that the models of the these two languages were more effective.

For development we used the following data sets:

- The second half of the development corpus provided by NIST for the LRE07 evaluation, divided into two sets for estimating the backend parameters.
- Half of the OHSU corpus provided by NIST for LRE05, halved again into two development sets.
- Development and test set provided by NIST for LRE03.
- About 6 hours of Farsi and 1 hour of Vietnamese excerpts from [19] have been selected to enrich the development set, which for these languages included segments that were (too) easily recognized.

5. Score combination

The five scores produced by the acoustic models (two gender dependent GMMs and PGMMs, and one gender independent GMM-SVM) and the two scores of the phonetic models are combined by means of a linear SVM backend to produce the final score. The final backend for the evaluation is trained on the set of scores obtained by the models on all the development data described in Section 4.

The scores of the segments of all the other languages/dialects in the close set are negative examples for training a target

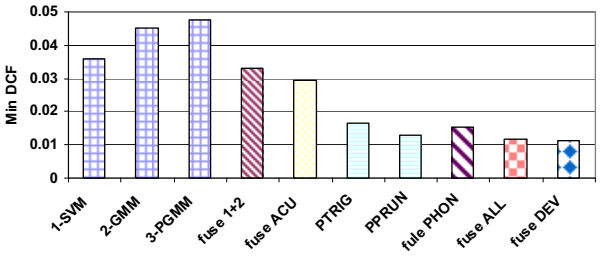

Figure 2. *Performance comparison of the acoustic and phonetic systems, alone and in combination.*

language SVM model. The score for the current language model is T-normalized by computing the statistics on the other language GMMs. The final score is then obtained by log-likelihood normalization [2][3].

According to the LRE07 evaluation plan [9], for each pair <test segment – language model> a decision True/False must be given in the result file, based on the final score produced by each model. Thus, a threshold must be evaluated on the development sets. Using one development subset for estimating the backend parameters, and testing on the other one, we evaluated for each task defined in [9] the threshold that optimized the cost performance. Inverting the role of the development subsets, another threshold is obtained. Our decision threshold is simply the mean of these two values.

6. Experiments

We present the results of a set of post-evaluation experiments aiming at analyzing the relative performance of the acoustic and phonetic systems and their combination. All the results refer to the close–set 30 sec General Language Recognition condition [9], where the task was to discriminate among 14 languages.

The first set of experiments evaluated the performance improvement due to the combination of different tokenizers. Figure 1 shows the range of the Decision Cost Functions (DCF) [9] obtained fusing a given number of tokenizers of different languages. It is worth noting that most of the gain is obtained fusing 3 or 4 tokenizers, and that the identity of the tokenizer languages is of minor importance when more than three of them are involved in the fusion. However, appreciable improvement is obtained by fusing more and more tokenizers.

The comparison of the performance of the acoustic and phonetic systems, alone and in combination, is summarized in Figure 2. The first three bars show the minimum DCF of the three acoustic systems alone. The next pair of bars gives the results of the combination of the GMM and GMM-SVM models, and their combination with the PGMM system respectively. The phonetic scores, shown in the next two bars of the figure, are far better than the acoustic ones for this condition. Moreover, the pruned multigram models are better than the trigram models, and as effective as the combination with the other systems. This does not happen for the shorter duration tests. For the shorter duration conditions, the PGMM system performs better than the ones based on GMMs. The last 2 bars represent the results of the combination of all the systems on the actual and development tests respectively.

Figure 3 shows the results of the fusion of the acoustic systems, of the phonetic systems, and their overall combination

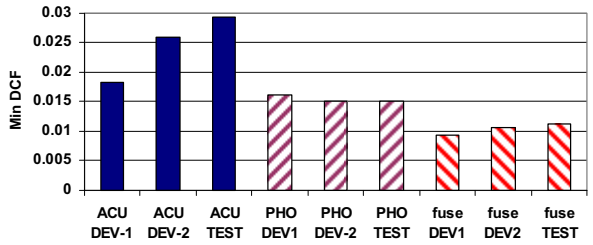

Figure 3. *Comparison of the performance of the systems on two development sets and on the LRE07 test set.*

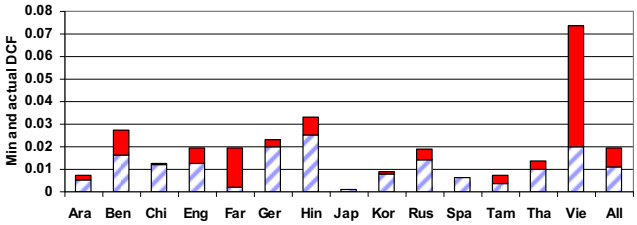

Figure 4. *Min (dashed) and actual DCF per language*

Table 2. *Min and actual (official) DCFs for the closed set tests we participated in LRE07.*

Languages	3s	10s	30s
General LR (***14 languages***)	**0.1331** **0.1436**	**0.0442** **0.0548**	**0.0112** **0.0195**
English (***American-Indian***)	**0.2031** **0.2172**	**0.1203** **0.1437**	**0.0813** **0.1094**
Chinese (***Cantonese***, ***Mandarin***, ***Min***,***Wu***)	**0.1699** **0.1878**	**0.0620** **0.0737**	**0.0217** **0.0304**
Mandarin (***Mainland***, ***Taiwan***)	**0.2577** **0.2672**	**0.1197** **0.1788**	**0.0891** **0.1135**
Hindustani (***Hindi***, ***Urdu***)	**0.3609** **0.3781**	**0.3406** **0.3984**	**0.3188** **0.3484**

using the LRE07 tests and the development sets described in Section 4.Comparing the first three bars, it can be noticed that the acoustic models evaluated on the two development sets did not generalize well to the LRE07 tests, whereas the phonetic models do not show such performance degradation (see the second set of bars in the figure)**.** The results of the fusion of the phonetic and acoustic systems, represented by the last set of three bars, shows that although the acoustic systems were not as good as the phonetic systems in this condition, their contribution is substantial to the overall performance.

Figure 4 summarizes the results obtained for the 14 languages of the LRE07 30 sec tests. Globally, a min DCF of 0.0110 has been obtained. The relevant gap in the average actual DCF of 0.0195 is mainly due to an erroneous setting of the decision threshold for Vietnamese. The results highlight the difficulties in setting appropriate decision thresholds whenever the training data of a language are scarce (see Bengali, Thai), or the test data are collected through previously unseen channels (see Farsi, Russian, and in particular Vietnamese).

Table 2 shows the min and actual DCFs obtained by our system in all the closed set tests of LRE07. Particularly interesting are the results for the Chinese languages considering that we did not use any oriental language phonetic transcriber.

7. Conclusions

The components and the performance of a system, exploiting state-of-the art techniques have been presented.
From the LRE07 evaluation we have learned that using data not homogeneous with the LRE corpora used by NIST in the evaluations is not the best choice. This leaves open the problem of language recognition in highly mismatched conditions.

8. References

[1] A.F. Martin, and A.N. Le, "NIST 2007 Language Recognition Evaluation", in Proc. Odyssey 2008,", 2008.

[2] W.M. Campbell, J.R. Campbell, D.A. Reynolds, E. Singer and P.A. Torres-Carrasquillo, "Support Vector Machines for Speaker and Language Recognition", in Computer Speech and Language, Vol. 20, pp. 210-229, 2006.

[3] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition", IEEE Trans. on Audio, Speech, and Language Processing. Vol. 15-7, pp. 1969-1978, 2007.

[4] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface and C. Vair, "Language Identification using Acoustic Models and Speaker Compensated Cepstral-Time Matrices", Proc. ICASSP 2007, Vol. IV, pp. 1013–1066, 2007.

[5] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Process. Lett., vol. 13, no. 5, pp. 308–311, May 2006.

[6] W.M. Campbell, J.R. Campbell, D.A. Reynolds, E. Singer and P.A. Torres-Carrasquillo, "Support Vector Machines for Speaker and Language Recognition", Computer Speech and Language, Vol. 20, pp. 210-229, 2006.

[7] L. Burget, P. Matejka, and J. Cernocky, "Discriminative Training Techniques for Acoustic Language Identification," in Proc. ICASSP 2006, Vol. I, pp. 209-212, 2006.

[8] F. Castaldo, E. Dalmasso, P. Laface , D. Colibro, C. Vair, "Acoustic Language Identification Using Fast Discriminative Training", Proc. Interspeech 2007, pp. 346-349, 2007.

[9] Avaliable at www.nist.gov/speech/tests/lang/2007/

[10] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, Vol. 10, pp. 19-41, 2000.

[11] P.A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features," in Proc. ICSLP 2002, pp. 90-93, 2002.

[12] Available at http://www.ldc.upenn.edu/Catalog.

[13] Available at http://www.speechdat.org/SpeechDat.html.

[14] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced Language Recognition using Cepstra and Phonotactics," IEEE Odyssey Speaker and Language Recognition Workshop, Puerto Rico, 2006.

[15] http://www.loquendo.com/en/technology/asr.htm

[16] W.M. Campbell, J.R.Campbell, D.A. Reynolds, D.A. Jones and T.R. Leek, "High-level Speaker Verification with Support Vector Machines", in Proc. ICASSP 2004, Vol I, pp. 73-76, 2004.

[17] W. M. Campbell, J. P. Campbell, T. P. Gleason, D. A. Reynolds, W, Shen, "Speaker Verification Using Support Vector Machines and High-Level Features", IEEE Trans. on Audio, Speech and Language Proc., vol. 15, n. 7, pp. 2085-2094, September 2007.

[18] S. Deligne, F. Bimbot, "Inference of variable-length linguistic and acoustic units by multigrams", Speech Communication vol. 23, pp. 223–241, 1997.

[19] /http://globalrecordings.net, "Telling the story of Jesus in every language", Global Recordings Network.