POLITECNICO DI TORINO
Repository ISTITUZIONALE

LOQUENDO-POLITECNICO DI TORINO SYSTEM FOR THE 2009 NIST LANGUAGE RECOGNITION EVALUATION

(Article begins on next page)

20 May 2024

LOQUENDO-POLITECNICO DI TORINO SYSTEM FOR THE 2009 NIST LANGUAGE RECOGNITION EVALUATION

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

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

{Fabio.Castaldo, Sandro.Cumani, Pietro.Laface}@polito.it

{Daniele.Colibro, Emanuele.Dalmasso, Claudio.Vair}@loquendo.com

ABSTRACT

This paper describes the system submitted by Loquendo and Politecnico di Torino (LPT) for the 2009 NIST Language Recognition Evaluation. The system is a combination of classifiers based on two core acoustic models and on two core 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 incremental training procedure that has been devised to deal with broadcast data, we also describe the models, the classification techniques that have been used for this evaluation, and we comment on the performance of the system components alone and in combination.

The system obtained using these techniques was among the best participants in this evaluation, obtaining on the 23 languages recognition task an actual DCFx100 of 1.6, 2.8, and 9.2 in the 30, 10 and 3 sec conditions respectively.

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

1. INTRODUCTION

A new challenge has been introduced in the 2009 NIST Language Recognition Evaluation (LRE) [1]: while all of the previous evaluation data consisted of Conversational Telephone Speech (CTS), two corpora of broadcast data consisting of Voice of America broadcasts in multiple languages have been distributed by NIST as additional data for this evaluation. Moreover, most of the 23 target languages were new, and several target languages had no CTS samples. For each of the target languages that were included in these corpora, a labeled development set was created by LDC including about 80 segments of approximately 30 seconds duration, audited by the provider (LDC) and found to contain narrowband speech in the target language.

Unfortunately this development corpus lacks the necessary intra-language variability due to channel, gender and speaker differences, to train robust language models. Thus, in the NIST evaluation plan it was allowed to collect additional training data from any publicly available source.

In this paper we first illustrate the incremental data selection and training procedure that has been devised to generate an appropriate development set for the narrowband speech in broadcast data. Then we describe the models that have been created, and the classification techniques that have been used for this evaluation. Finally, we comment on the performance of each system component and of its combination with the others, highlighting some still open problems such as the poor results obtained testing CTS data using models trained with narrowband speech collected from broadcast corpora.

2. TRAINING AND DEVELOPMENT DATA

While most of the CTS data were available from previous NIST evaluations, narrowband speech segments from broadcast data (*broadcast* for short in the following) had to be carefully selected to create the language models according to an incremental procedure starting from the Voice of America corpora provided by NIST for the 2009 evaluation, which contain speech in most of the 23 target languages [1]. These corpora, referred to as VOA2 and VOA3 in the following, were supplied down-sampled to 8 KHz, in 8-bit mu-law format. VOA3 programs have VOA supplied language labels, while those from VOA2 have associated a set of hypothesized language labels created by an automatic procedure [2] , which may be erroneous.

2.1. Telephone development data

The following CTS corpora were used for training:
- The Callfriend corpus [3]. The conversations in this corpus were split into slices of approximately 150s.
- The corpora provided by NIST for LRE03, LRE05 and LRE07.
- The Russian through switched telephone network [3].
- The Cantonese and Portuguese data in the 22 Language OGI corpus [4].

2.2. Broadcast development data

The development corpora were incrementally created to include as far as possible the intra-language variability due to channel, gender and speaker differences. To obtain a language recognition system with good generalization capabilities, we had to generate a development set - further split in training, calibration and test subsets - covering the mentioned variability with a sufficient amount of examples, and without speaker overlap among the subsets. The LRE2009 broadcast development data and the audited corpus provided by NIST did not satisfy the previous requirements for the following reasons:
- The segments are often from the same speaker, as detected by our speaker recognizer [5], and confirmed by the "uniq_spkr" field of the audited data set to 'False'.
- After filtering the same speaker segments, a small number of segments remain for some languages.
- The speaker genders within a language are not balanced.
- Excluding "French", the segments of all the other languages are either telephone or broadcast.

- No audited data were available for Hindi, Russian, Spanish and Urdu on VOA3, only the automatic segmentation from Brno University (BUT) was given.
- No segmentation was provided in the first release of the VOA3 development data for Cantonese, Korean, Mandarin, and Vietnamese.

For these 8 missing languages only the automatic language hypotheses provided by BUT were available for VOA2 data.

Overall, we had only CTS data for 13 languages, only broadcast data for 21 languages, and data in both conditions for 11 of these languages. The "Persian" broadcast data in VOA2 and VOA3 were considered samples of the Farsi language.

2.3. Additional checked development data

For the 8 languages lacking audited broadcast, segments have been generated accessing the VOA site [6] looking for the original MP3 files, also included – down-sampled – in the VOA3 disk. The goal was to collect about 300 broadcast segments per language, which were first processed by detecting narrowband fragments with a procedure similar to the one described in [2]. The candidates were checked to eliminate segments including music, bad channel distortions, and fragments of other languages that were evidently not corresponding to the file labels.

The telephone and the audited broadcast data, plus the data of this additional set were evenly split into a train and a test set. Due to the scarcity of data we did not create a calibration set, performing simple score normalization by ZT-norm.

The segments belonging to the same speaker were included in the same set. The speaker information was obtained by running our speaker recognizer on the broadcast segments.

This set was used to train preliminary "bootstrap" models, one acoustic and one phonetic. The tests performed using these bootstrap models have highlighted the aforementioned problems related to the adequacy of the development data, and the necessity of further enriching the development sets.

2.4. Additional not audited development data

The samples necessary to enrich the development sets with new speakers and more segments were selected from the VOA3 and VOA2 data. For the VOA3 database we assume that the file label correctly identify the corresponding language. Among the VOA3 data segmented by BUT, but not audited, we selected those allowing us to include new speakers in the train, calibration and test sets. The selection of speakers for each language was performed by means of the speaker recognizer. The audited segments were processed first, followed by the checked ones, and finally by the others in order to discard segments belonging to frequently appearing speakers. Whenever the best recognition score obtained by a segment was less than a predefined threshold, a new speaker model was added to the current set of speaker models.

The selection of segments from the VOA2 database was more complex, and possibly error prone. Language recognition was performed on each segment using a system combining the bootstrap models. A segment was selected only if it had associated a score greater than a given rather high threshold, and if the 1-best language hypothesis of our system matched the 1-best hypothesis provided by the BUT system. The speaker selection procedure was applied also to these segments. The number of different speakers per language resulting from this procedure is about 45 on average.

Table 1: Number of segments and total segment time selected from the Voice of America broadcast corpora

Set	Time	Broadcast Corpora					
		voa3_A	voa2_A	ftp_C	voa3_U	voa2_U	ftp_U
Train	40 h	529	116	316	1955	590	66
Extended train	48 h	114	22	65	2483	574	151
Calibration and Test	34 h	396	85	329	1866	449	45

Table 1 shows the number of segments of broadcast data included in the final sets, and their total duration. In these tables the *ftp* label refers to the narrowband segments extracted from the original MP3 files available in the VOA site. Suffixes *A*, *C* and *U* refer to audited, checked, and additional unchecked segments respectively.

3. LANGUAGE IDENTIFICATION SYSTEM

The LPT system is the combination of classifiers based on two acoustic core models and two core phonetic tokenizers.

The acoustic models are Gaussian Mixture Models obtained from a common Universal Background Model (UBM). The UBM and the language GMMs consist of mixtures of 2048 Gaussians. The observation vector includes the usual 56 parameters: the first 7 Mel frequency cepstral coefficients and their 7-1-3-7 Shifted Delta (SDC) coefficients.

Two core acoustic models have been trained, both based on Gaussian Mixtures. These models will be referred to in the following as *pushed* GMMs and MMIE trained GMMs respectively.

3.1.1. Pushed GMMs

These discriminative models are obtained by a combination of GMMs through the information provided by Support Vector Machine (SVM) classifiers (GMM-SVM) according to the method proposed in [7]. A model per utterance is obtained by using Maximum A Posteriori (MAP) adaptation with a small relevance factor. Channel dependent but gender independent GMMs have been trained to avoid reducing the number of positive class examples. Training the channel dependent model of a target language is performed using as positive class the set of the channel dependent GMMs of that language, and all the GMMs of the competitor languages as negative classes (irrespective of channel).

The total number of models that we use for scoring an unknown segment with this system is 34: the channel dependent models are 22 (11 CTS and 11 broadcast) and the single channel models are 12 (2 telephone and 10 broadcast models only).

3.1.2. MMIE trained GMMs

The second set of acoustic models is trained by Maximum Mutual Information Estimation (MMIE) [8].

Training is performed on frame blocks separated by silences, identified by a recognizer of broad phonetic classes. Gender dependent models were trained with 7 iterations, bootstrapped from gender independent pushed GMMs. The gender information was provided by labels, when available, or by our speaker recognizer trained to perform only gender detection.

Since we use channel independent but gender dependent models, the number of scores per segment is 46, 23 per gender.

3.1.3. Nuisance compensation

For both models, the features domain compensation approach that was successful in the previous evaluation [9] was applied to reduce channel and speaker variability within the same language. We estimated a subspace that represents the distortions due to inter-language variability, and compensate these distortions in the domain of the features using factor analysis [10]. The subspace of the intra-language variability is modeled by a low rank matrix U, of dimension 120 in these experiments. The U matrix and UBM that were trained for the LRE07 evaluation, with telephone data only, have been used to obtain the GMM bootstrap models. When the training set was enriched as illustrated in the Section 2, new matrices U and UBMs have been estimated. Each new U matrix is estimated by collecting the differences between GMM supervectors of each language. These differences have been performed separately for segments labeled as broadcast or telephone and among broadcast and telephone segments.

3.2. Phone models

The combination of acoustic with phonetic systems has been successful in the past evaluations [9][11]. In particular, in LRE07 we exploited the availability of several languages in the Loquendo-ASR recognizer [12] to implement a phonetic system based on the Parallel Phone tokenizer-SVM [13].

3.2.1. 1-best LID SVM

The first phonetic system is based on the standard Loquendo-ASR decoder, which uses hybrid ANN-HMM models described in [14]. The decoder uses a phone-loop grammar with diphone transition constraints, and produces the 1-best phone strings for each segment. For this system, 12 different phone grammars have been used in parallel to collect the statistics of the n-gram phone occurrences in each segment for the following languages: French, German, Greek, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Turkish, UK and US English.

From each phone sequence produced by one of our phonetic transcribers on the same segment, the frequency of occurrence of each n-gram is computed and 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 produce the so called Term Frequency Log-Likelihood Ratio (TFLLR) kernel [15] used in the SVM approach to language identification [7][9].

Channel dependent linear SVM models of the target languages were trained. Two different TFLLR kernels have been used, the first one based on 3-grams, and the second one relying on pruned n-grams of order higher than 3 [8].

3.2.2 Lattice 3-grams

The second phonetic system is based on the same features and phone-loop grammars, but uses slightly different ANN acoustic models and a search engine that produces phone lattices. The number of language transcribers for this system is 10 (a new language, Catalan, is included in the previous list of languages, whereas Greek, Portuguese and UK English were excluded).
Again, channel dependent SVM models were trained computing the 3-gram probability using the expected counts from a lattice rather than the statistics from the 1-best sequence [16] [17].

6. SCORE NORMALIZATION AND COMBINATION

The system produces its final scores by combining the scores of the 5 sub-systems illustrated in Section 3. Since the dimension of

Table 2: Performance of the 5 sub-system on the 30s development set (minDCFx100) and on the evaluation sets (actual DCFx100)

TEST ON	SYSTEMS					
	Pushed GMMs	MMIE GMMs	3-grams	Multi-grams	Lattice	Fusion
Development	1.48	1.70	1.09	1.12	1.06	0.86
Evaluation	2.13	2.15	1.64	1.53	1.47	1.16
Broadcast	2.03	2.01	1.63	1.51	1.39	1.08
Telephone	3.09	3.47	2.25	2.26	2.49	2.06

the score vectors for all the channel dependent sub-systems is 34, whereas it is 23+23 for the MMIE GMMs sub-system, the total number of scores is 182.

The back-end training procedure follows the normalization and calibration procedure proposed in [11][18] and uses the FoCal multiclass toolkit [19].

The final back-ends for the evaluation were trained on the set of scores obtained by the models on all the development and test data described in Section 2. Separate back-ends were trained for the 3, 10, and 30 sec conditions using the development subsets of the corresponding durations.

For each sub-system a set of channel dependent Gaussian back-ends has been trained. In particular, the space of the scores produced by each sub-system is transformed by means of LDA, and 34 Gaussians with common full covariance are trained by Maximum Likelihood Estimation.

The output of each backend is a vector of 34 scores, 22 of them related to languages having both telephone and broadcast development data, 10 to languages having only broadcast data, and the remaining two – American and Indian English – having only telephone data.

The raw score vectors are transformed into log likelihood vectors by applying the Gaussian back-ends, and the calibrated fusion of the 5 sub-systems is performed by means of multiclass Linear Logistic Regression (LLR), which finds the transformation parameters that optimize the multi-class Cllr objective function [18].

The best log likelihood is selected for the 11 languages having both broadcast and telephone scores.

These probabilities are transformed into the log likelihood ratio score llr=log P(segment|Language)/P(segment|¬Language) using the a priori probabilities and costs given in the NIST LRE09 evaluation plan [1], thus the decision threshold is simply 0.

7. RESULTS

The five subsystems and their fusion were assessed on a telephone LRE07 subset (restricted to the LRE09 target languages) and on the broadcast development data of Section 2. The development set was further split into 2 subsets, used for calibration and testing purposes. Both subsets were used for calibration and testing, exchanging their roles.
The first row of Table 2 summarizes the performance of the 5 sub-system on 30 sec segments of the LRE09 development data, in terms of minDCFx100 [1]. The second row gives the actual DCFx100 obtained on the evaluation set. The last two rows show the performance of the sub-systems on broadcast and telephone test data only. The results of the fusion of the sub-system are given in the last column.

We can notice different behaviours of the subsystems on different subsets. For instance, MMIE GMMs and lattice models are better on broadcast data, whereas pushed GMMs and 1-best 3-gram systems perform better on CTS.

Another comparison of the sub-systems can be appreciated looking at the bar chart of Figure 1, where the minimum and actual DCF obtained on the 30 sec evaluation set are shown. The first three bars are related to the two acoustic subsystems and their fusion. The next four bars are related to the phonetic subsystems and their fusion. The gap between the acoustic and phonetic models is about 27%, far smaller then the 50% gap we had in the LRE07 evaluation. This improvement was probably due to the use of pushed models and of better MMIE models obtained starting from pushed models. The fusion of acoustic and phonetic models is rather effective, with a 17% of relative minimum DCF reduction. Looking at the last bar, which shows the performance of the sub-system combination, a rather high (30% relative) calibration error, given by the difference between actual and minimum DCF, still remain.

It is interesting noting that, whereas for the 30 sec test condition the best combination of 5/6 decoders almost reaches the accuracy obtained with 12 phone recognizers, for the 3 second condition using all the transcribers improves the minimum DCF from 0.127 to 0.116.

Finally, the false alarm and the miss rates obtained using models trained with CTS only and CTS plus broadcast data show that these models perform quite well even when tested on broadcast segments. Unfortunately, the reverse is not true: broadcast models do not perform well on CTS data. This is valid for every language, and our results confirm the findings in [2].

8. CONCLUSIONS

An incremental training procedure has been presented exploiting state-of-the art techniques in speaker and in language recognition to select and label narrowband segments within broadcast data.

Although very good results have been obtained in the LRE09 evaluation, using discriminative channel compensated acoustic models and several phonetic transcribers, our results confirm that still open problems remain in using models trained with easily available broadcast data for recognizing CTS data. Different speaking styles - characterized by good pronunciation - and high mismatch in channel characteristics seem to weaken the models against conversational speech.

9. REFERENCES

[1] Avaliable at http://www.itl.nist.gov/iad/mig//tests/lang/2009

[2] O. Plchot, V. Hubeika, L. Burget, P. Schwarz, P. Matejka, J. Cernocky, "Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition", Technical Report http://www.nist.gov/speech/tests/lre/2009/radio_broadcasts.pdf

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

[4] Available at https://www.cslu.ogi.edu/corpora/22lang.

[5] E. Dalmasso, F. Castaldo, P. Laface, D. Colibro, C. Vair, "Loquendo-Politecnico di Torino's 2008 NIST Speaker Recognition Evaluation System", ICASSP 2009, pp. 4313–4216, 2009.

[6] Available at ftp://8475.ftp.storage.akadns.net/mp3/voa

[7] W. M. Campbell, D. E. Sturim, P. Torres-Carrasquillo, D. A. Reynolds, "A Comparison of Subspace Feature-Domain Methods for Language Recognition", Interspeech 2008, pp-309-312, 2008.

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

[9] F. Castaldo, E. Dalmasso, P. Laface, D. Colibro, C. Vair, "Politecnico di Torino System for the 2007 NIST Language Recognition Evaluation", Interspeech-2008, pp-297-300, 2008.

[10] 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.

[11] P.Matejka, L. Burget, O. Glembek, P. Schwarz, V. Hubeika, M. Fapso, T. Mikolov, O. Plchot, J. Cernocky, "BUT language recognition system for NIST 2007 evaluations", Interspeech 2007, pp. 739-742.

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

[13] 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.

[14] D. Albesano, R. Gemello, F. Mana, "Hybrid HMM-NN for Speech recognition and Prior Class Probabilities",International Conference on Neural Information Processing, vol. 5, pp.2391-2394, 2002.

[15] 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.

[16] J.L. Gauvain, A. Messaoudi H. Schwenk, "Language Recognition using Phone Lattices", ICSLP-2004, pp.1283-1286, 2004.

[17] W.M. Campbell, F. Richardson, D. A. Reynolds, "Language recognition with word lattices and support vector machines", ICASSP 2007, pp. 989 –992, 2007.

[18] N. Brummer, D. Van Leeuwen. "On Calibration of Language Recognition Scores", Proc. 2006 IEEE Odyssey – The Speaker and Language Recognition Workshop, pp. 1-8, 2006.

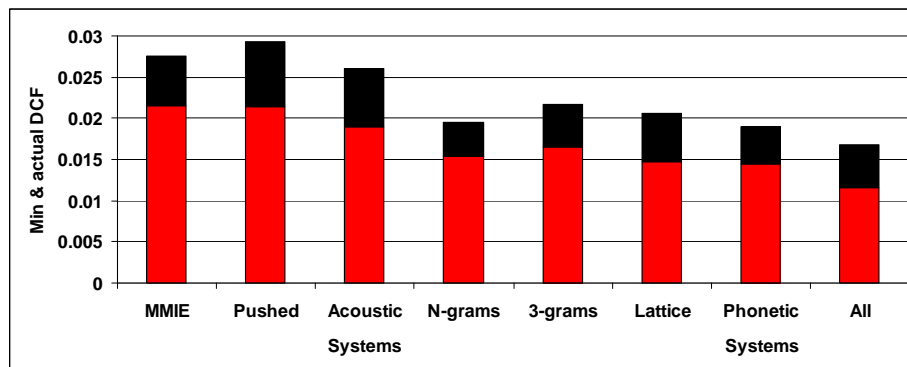[19] Avaliable at niko.brummer.googlepages.com/focalmulticlass

Figure 1. Comparison of the systems in terms of the minimum and actual DCF for the LRE09 closed-set 30s tests