THE L^{2}F LANGUAGE VERIFICATION SYSTEMS FOR ALBAYZIN-08 EVALUATION

Alberto Abad and Isabel Trancoso

L^{2}F - Spoken Language Systems Lab
INESC-ID / IST, Lisboa, Portugal
{Alberto.Abad,Isabel.Trancoso}@l2f.inesc-id.pt

RESUMEN

This paper presents a description of the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) Language Verification systems submitted to the ALBAYZIN-08 evaluation. Two completely different systems are presented for the restricted and the unrestricted evaluation. The restricted system relies on Gaussian mixtures models to classify language using the acoustic characteristics of the speech signals extracted by a front-end of shifted deltas. The unrestricted system is a Parallel Phone Recognition and Language Modeling system based on four different phone tokenizers. Results on the development data set for the different systems and evaluation conditions are presented. Additionally, measurements of the computational cost of processing the evaluation data set are provided.

1. INTRODUCTION

The "Red Temática en Tecnologías del Habla" (RTTH) has organized in the recent years a series of evaluations - so called ALBAYZIN evaluations - in some relevant speech processing topics devoted to encourage language research activities on the four official languages of Spain: Castilian, Catalan, Basque and Galician.

Similar to the well-known NIST Language Recognition Evaluation series, a Language Verification (LV) task has been proposed in ALBAYZIN-08. The objective is to determinate if each one of the four official languages of Spain is spoken (or not) in a given test file.

Language verification and recognition approaches can be classified according to the kind of source of information that they rely on. Most successful systems are based on the exploitation of the acoustic phonetics, that is the acoustic characteristics of each language, or the phonotactics which are the rules that govern the phone combinations in a language.

Acoustic systems model each language short-term acoustics by means of stochastic models/classifiers such as Gaussian mixtures models (GMM), Neural Networks (NN) or Support Vector Machines (SVM). Phonotactic systems usually use language dependent stochastic grammars to model phonemes or broad categories of phonemes extracted by a tokenizer.

This paper presents the LV systems developed by the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) for the ALBAYZIN-08 campaign. In accordance with the evaluation conditions, two different systems have been presented: an acoustic system based on GMM modeling for the restricted evaluation (GMM-LV restricted), and a phonotactic Parallel Phone Recognition and Language Modeling system for the unrestricted evaluation (PPRLM-LV unrestricted). The next Section 2 presents a brief description of the task, the data provided for the evaluation and the evaluation metrics. Sections 3 and 4 describe the GMM-LV restricted and the PPRLM-LV unrestricted systems, respectively. Measurements of the computational deployment in the processing of the evaluation data set are also provided. In Section 5 results obtained by the two systems in the different evaluation conditions with the development data set are presented. Finally, Section 6 presents our main conclusions.

2. ALBAYZIN-08 LV: TASK, DATA AND METRIC DESCRIPTION

Detailed information on the ALBAYZIN-08 LV campaign can be found in the evaluation plan document[1].

2.1. Task and evaluation conditions

The task consists of deciding whether a speech segment belongs to each one of the four target languages (Castilian, Catalan, Basque and Galician) or not. That is, for each test signal, four decision results (true or false) for each one of the target languages are produced, together with a score of the decision.

Two system evaluation categories are proposed: one for restricted systems that rely only on the data provided for the evaluation, and another for unrestricted systems which can use any data or incorporate subsystems that have been trained with external data, for instance phone classifiers or voice activity detectors.

Additionally, the systems can be evaluated in closed mode or open mode. In contrast to the closed mode, in the open mode speech segments from unknown languages different from the target ones can appear in the test data

This work was partially funded by the FCT project PoSTPort (PTDC/PLP/72404/2006) and by the European project Vidi-Video.

and are taken into account for the systems performance evaluation.

2.2. Train, development and test data

All the data provided for the ALBAYZIN-08 evaluation are TV programs captured at 16 kHz. The training data set consists of approximately 8 hours per target language, in several files of varying length. The test data set consists of 1800 files with speech of the four target languages and in other unknown languages of 3 different durations: 3, 10 and 30 seconds. Additionally, a development data set consisting of also 1800 files of similar characteristics to the evaluation test set was provided with language identification labels.

2.3. Performance metric

An average performance score based on the false positive and false alarm rates obtained by the evaluating systems is used. The performance score, hereinafter referred to as C_{avg}, is computed independently for each test length duration (3, 10 and 30 seconds). Further details about the metrics can be found in [1].

3. THE GMM-LV RESTRICTED SYSTEM

For the restricted system a GMM acoustic modeling approach was considered to be the most adequate, since it does not need phonetic or word-level transcriptions of any kind. In this section, a detailed description of the developed system is provided.

3.1. Training data splitting

The training data provided for each language (~ 8 hours) was split into two distinct sets: the *models data* of approximately 400 minutes per language was used for training the acoustic models, and the *back-end data* of approximately 100 minutes per language was used for normalizing scores and back-end development.

3.2. System description

The system is a conventional GMM classifier. For each target language, a GMM is trained with the *model data* of that language extracted by the selected front-end. During test, the acoustic language models are used to compute the log-likelihood scores of a given speech signal for each language. These likelihood scores of the four language models are then processed to produce a decision for each target language. Only one GMM classifier for one set of features has been trained. Figure 1 shows a diagram block of the GMM-LV restricted system.

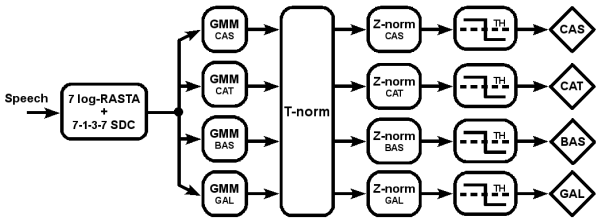

Figure 1. *Block diagram of the GMM-LV restricted system presented for ALBAYZIN-08 evaluation.*

3.2.1. Feature extraction

The extracted features are Perpetual Linear Prediction static features with log-RelAtive SpecTrAl speech processing (log-RASTA), and a stacked vector of shifted delta cepstra (SDC) of the same log-RASTA features. Concretely, 7 log-RASTA static features and a 7-1-3-7 SDC parameter configuration are computed, resulting in a final feature vector of 56 components.

On the one hand, log-RASTA features are known to be a robust representation for speech processing applications [2]. On the other hand, it has been shown that the use of SDC features (created by stacking delta cepstra computed across several frames) allows improved performances in LV tasks [3]. The selected front-end showed remarkable improvements compared to other evaluated feature representations during the development of the systems, such as mel-frequency cepstral coefficients, perceptual linear prediction features or the advanced ETSI front-end features.

3.2.2. Acoustic modelling

Gaussian mixture models of 1024 mixtures were trained for each target language using the *models data*. Each acoustic model was first initialized by means of vector quantization estimation. Then, 10 maximum likelihood Estimation Maximization iterations were applied to obtain the final language models.

More sophisticated model training procedures were also tested, without achieving significant differences. In particular, the use of a universal background model (UBM) for Bayesian adaptation to the target languages. This method was finally discarded, and UBM has not been used neither for model adaptation nor for score normalization.

3.2.3. Back-end: normalization and scoring

For each test utterance, log-likelihoods of the four acoustic models (Castilian, Catalan, Basque and Galician) were obtained. The log-likelihood of the claimed or tested language was T-normalized with the mean of the log-likelihoods of the other three competing languages.

The *back-end data* was split in shorter segments, according to an energy-based speech detector segmentation

system, in order to be more similar to the test data. T-norm scores of this *back-end data* were computed. Then, mean and variance of the T-norm score of a concrete language was estimated with the *back-end data* sets of the competing languages. During test, the mean and variance estimated for a target language were used to apply Z-norm to the T-normalized log-likelihood of this language, in order to obtain the final score used for decision.

The final decision threshold was selected in order to deploy a balanced performance (close to minimum C_{avg}) for the 3, 10 and 30 seconds evaluation conditions. It is worth noticing that a unique threshold is used independently on the claimed target language.

The difference between the closed and open systems for the restricted evaluation is on the decision threshold selected, that is slightly more selective in the open system to reduce the increased false alarm rate due to presence of unknown language speech sources.

3.3. Processing time

All the experiments in this paper were run in an Intel Quadcore 2.4 GHz (Q6600) machine with 8 GBytes of DDR2 RAM at 667 Mhz.

The total time deployed by the restricted system (both closed and open) in performing the test evaluation was approximately 19 minutes. Since the evaluation test set has a total duration of around 458 minutes, this result corresponds approximately to 0.04xRT.

4. THE PPRLM-LV UNRESTRICTED SYSTEM

The unrestricted system presented for the ALBAYZIN-08 LV evaluation is a PPRLM system that exploits the phonotactic information extracted by four parallel tokenizers: European Portuguese, Brazilian Portuguese, European Spanish (Castilian) and American English. The key aspect of this type of systems is the need for robust phonetic classifiers that generally need to be trained with word-level or phonetic level transcriptions. At the INESC-ID's L^2F group we have been working for several years in Large Vocabulary Continuous Speech Recognition (LVCSR) using hybrid Artificial Neural Network Hidden Markov models (ANN/HMM) recognizers, the so-called connectionist paradigm. During the last years, we have been developing phonetic classifiers (Multi-layer Perceptrons, MLP) for our current recognition systems in several languages. In this section, we present a detailed description of the PPRLM-LV system and its several components.

4.1. Training data splitting

Like in the case of the GMM-LV system, the training data has been split into two sets, one for training stochastic language models (*models data* of approximately 400 minutes per language), and the other for back-end development (*back-end data* of approximately 100 minutes per language).

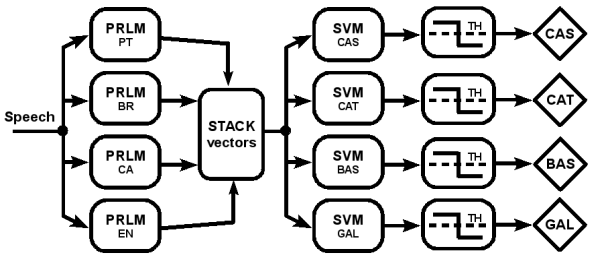

Figure 2. *Block diagram of the PPRLM-LV unrestricted system presented for ALBAYZIN-08 evaluation.*

The four phonetic classifiers used by the PPRLM-LV system were trained with additional external data. For the European Portuguese classifier, 57 hours of manually annotated data and more than 300 hours of automatically transcribed broadcast news (BN) data were used. The Brazilian Portuguese classifier was trained with around 13 hours of BN data. The Spanish system used 14 hours of manually annotated data and 78 hours of automatically transcribed data. Finally, the English system was developed with the HUB-4 96 and HUB-4 97 data sets, that contain around 142 hours of TV and Radio Broadcast data.

4.2. System description

The PPRLM-LV unrestricted system first uses the four phonetic tokenizers to extract the phonetic sequence of the *model data* of each target language. Then, for each target language and for each tokenizer a different phonotactic n-gram language model is trained. During test, the phonetic sequence of a given speech signal is extracted with the phonetic classifiers and the likelihood of each target language model is evaluated, resulting in a total of 16 likelihood scores (4 target languages x 4 phonetic tokenizers). These likelihood scores are normalized and combined with a Support Vector Machine approach for obtaining a final identification and probability score per target language. Figure 2 shows a diagram block of the PPRLM-LV unrestricted system.

4.2.1. Phonetic tokenizers/classifiers

The tokenization of the input speech data in both training and testing is done with the neural networks that are part of our hybrid Automatic Speech Recognition (ASR) system named AUDIMUS. This kind of recognizers are generally composed by one or more phoneme classification networks, particularly MultiLayer Perceptrons (MLP), that estimate the posterior probabilities of the different phonemes for a given input speech frame (and its context). Concretely, the system combines three MLP outputs trained with Perceptual Linear Prediction features (13 static + first derivative), log-RelAtive SpecTrAl features (13 static

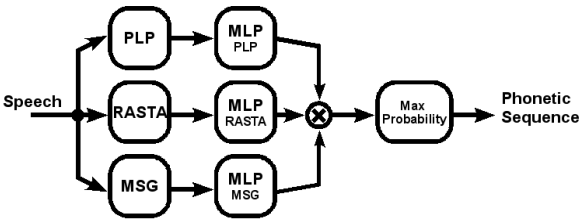

Figure 3. *Block diagram of a multi-stream MLP phonetic classifier used in the European Portuguese, Brazilian Portuguese, Castilian and English PRLM systems.*

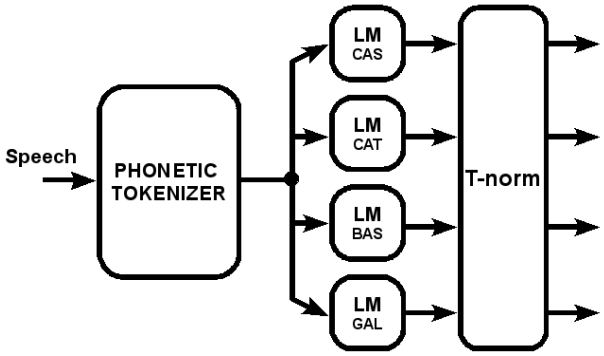

Figure 4. *Block diagram of one PRLM language verification system.*

+ first derivative) and Modulation SpectroGram features (28 static). Figure 3 shows the structure of one of the multi-stream phonetic classifiers used in this work. A detailed description of the European Portuguese BN transcription system can be found in [4].

The size of the neural networks of each ASR system (European Portuguese, Brazilian Portuguese, European Spanish and American English) differs due to the different amounts of training data. However, it is worth noticing the differences in the output layer, that is, the number of different phonetic tokens to classify. In the case of European Portuguese, 39 phonetic tokens (complete Portuguese phone set + silence) are considered. In the Brazilian, Spanish and English recognizers, besides the complete phone set of each language plus silence, additional sub-phonetic units are also classified. These units are mainly phoneme regions (transitional left and right regions and steady phone nucleus) and diphone units [5].

4.2.2. Phonotactics modeling and normalization

For every phonetic or sub-phonetic tokenizer, the phonotactics of each target language are modelled with a 3-gram model. For that purpose the SRILM toolkit has been used [6].

During test, a vector with the four likelihoods obtained with four competing target language models is formed for every tokenizer. Similarly to what was done for the GMM-LV system, we decided not to use background models for score normalization. Instead of background normalization, T-norm of the mean likelihood of the three competing languages was applied to the score of a claimed language. A diagram of one single phonotactic system is shown in Figure 4.

4.2.3. Linear SVM combination

In order to combine the 4-element (one per target language) T-normalized vectors obtained from each independent phone tokenizer, linear Support Vector Machines are trained with the libSVM toolkit [7].

The *back-end data* portion was first segmented into shorter segments by a speech-non-speech detector, and the four 4-element T-normalized vectors scores were extracted and stacked to form a single 16-element vector. Then, four binary "1 versus all" classifiers were trained for every target language in order to obtain probability estimations. In fact, due to the high confusability between Castilian and Galician observed during the development of the system, it was decided to use a two step classification procedure when claimed languages were Castilian or Galician. A "2 versus all" classifier is trained to detect Castilian and Galician and then a "1 versus 1" classifier is used to disambiguate between these two. It is worth noticing that the SVM classifiers were used to estimate the probability of each target languages and that decisions were finally taken based on this probability score and the decision threshold selected.

Like in the GMM-LV system, the decision threshold was adjusted to obtain the best possible performance for the several evaluation conditions. Again, a different threshold is selected for the closed and open evaluation modes.

4.3. Processing time

Using the above mentioned machine, the total time deployed by the unrestricted system (both closed and open) was approximately 148 minutes, corresponding to approximately 0.32xRT. It is worth to notice that the time consumed on loading the phonetic networks of each one of the PRLM systems is included in this time computation and that the networks are loaded for each testing file.

5. RESULTS ON THE DEVELOPMENT SET

Table 1 presents the results obtained in the development set, for the two systems in all the described conditions. The results confirmed our expectations. The best ones were obtained with the unrestricted system. The open mode is significantly more challenging than the closed

one. The use of longer segments contributes to a smaller error rate.

System	Condition	30 sec	10 sec	3 sec
Restricted	Closed	0.1556	0.1986	0.2462
Restricted	Open	0.1952	0.2221	0.2648
Unrestricted	Closed	0.0281	0.0663	0.1635
Unrestricted	Open	0.0838	0.1148	0.1969

Table 1. *C_{avg} performance on the ALBAYZIN-08 LV development set of the GMM-LV restricted and the PPRLM-LV unrestricted systems in both open an closed mode.*

6. SUMMARY AND CONCLUSIONS

The experiments described in this paper using both a restricted system based on GMM models and an unrestricted system based on phonotactic models confirmed the advantages of using extra knowledge sources in the language verification task. It would be interesting to add to the dataset the other language spoken in the Iberian Peninsula (European Portuguese), as well as Brazilian Portuguese and the different Latin American Spanish varieties, and detect the confusability between all the languages.

7. BIBLIOGRAPHY

[1] "Plan de Evaluación de Sistemas ALBAYZIN-08 Verificación de la Lengua (ALBAYZIN-08 VL)", URL: http://jth2008.ehu.es/Plan_Albayzin-08_VL_final.pdf.

[2] Hermansky, H. and Morgan, N., "RASTA processing of speech", IEEE Transactions on Speech and Audio Processing, Vol. 2(4), pp 578-589, Oct 1994.

[3] Torres-Carrasquillo, P. A. et alt., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", in Proceedings of ICSLP 2002, pp. 89-92, Denver, Colorado, Sep 2002.

[4] Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I., "Audimus.media: a broadcast news speech recognition system for the european portuguese language", in Proc. PROPOR 2003, Faro, Portugal, 2003.

[5] Abad, A. and Neto, J., "Incorporating Acoustical Modelling of Phone Transitions in an Hybrid ANN/HMM Speech Recognizer", in Proc. INTERSPEECH-08, Brisbane, Australia, Sep 2008.

[6] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", in Proc. ICSLP 2002, Denver, Colorado, Sep 2002.

[7] Chang, C.-C. and Lin, C-J, "LIBSVM - A Library for Support Vector Machines", URL: http://www.csie.ntu.edu.tw/ cjlin/libsvm/index.html.