Transcription of Multi-variety Portuguese Media Contents

Alberto Abad[1], Hugo Meinedo[1], Isabel Trancoso[1,2], and João Neto[1,2]

[1] INESC-ID Lisboa, Portugal
[2] Instituto Superior Técnico, Lisboa, Portugal
alberto.abad,hugo.meinedo,
isabel.trancoso,joao.neto@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt/

Abstract. Current automatic transcription technology applied to media contents is an important medium that not only allows generating subtitles, but also enables data search and retrieval capabilities over multimedia streams. Among others, one of the most important challenges that transcription systems have to deal with is speaker accent variability. In this work we study the importance of accent variability for three broad varieties of Portuguese: African Portuguese, Brazilian Portuguese and European Portuguese. Then, we propose a multi-variety transcription system based on the combination of variety identification followed by specific variety-dependent transcription systems.

Keywords: automatic speech recognition, speaker accent variability, accent identification, broadcast news transcription.

1 Introduction

The presence of speech of different accents of the same language and the need for robustly dealing with them poses a significant challenge in automatic speech recognition (ASR) systems [1]. Thus, the influence of accent as a speaker variability factor in speech recognition has been the focus of intense research both for dialectal and foreign/non-native speech. In general, approaches to deal with accented speech can be coarsely classified into two groups [2]. First, those based on the adaptation of a general non-accented speech recognizer to the particular characteristics of one speaker or a small group of speakers. Typically, in these approaches acoustic models [3] and pronunciation lexicons [4] are adapted with a small amount of data. A second possible approach when a large amount of training data for each accent is available consists of building complete accent or variety specific speech recognition systems [2]. The key point in this type of approach is the need for an automatic method to identify and select the appropriate system for each accent.

Besides applications such as voice operated interactive systems, the need for robust recognition of several language varieties can be also crucial in automatic

H. Caseli et al. (Eds.): PROPOR 2012, LNAI 7243, pp. 409–420, 2012.

media contents transcription. Let us take, for instance, the example of a news subscription service that crawls for different media sources to automatically transcribe and then provide results to the users according to their preferences. A certain user may be interested only in news in a specific language, but he/she may be indifferent to the specific variety, the variety information may not be available or even the media content may present a mixture of speech of different varieties. In any of these cases, it is necessary that the media transcriber system reacts robustly to this type of variability.

At the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) we have developed in the last years a media content processing system that integrates several speech and language technologies for European Portuguese. Among other applications, this system has been used as the core of the fully automatic speech recognition subtitling system that is running on the main news shows of the public TV channel in Portugal (RTP), since March 2008 [5]. In fact, although Portuguese is the seventh most spoken language in the world with around 178 Million L1-speakers [6], only about five percent speak Portuguese with an European Portuguese (EP) accent. Consequently, one can expect that a considerable amount of media content is generated everyday in other Portuguese varieties. This fact motivated the development of variety dependent recognition modules for two other broad varieties of Portuguese: Brazilian Portuguese (BP) [7] and African Portuguese (AP) [8]. BP is the variety spoken in South America and it is the one with the largest number of speakers. AP encompasses the varieties spoken in one of the five PALOP countries (African Countries with Portuguese as Official Language): Angola, Cape Verde, Guinea-Bissau, Mozambique and São Tomé and Príncipe. The motivation to consider AP as a broad geographical variety is related to the difficulty to obtain data from each individual country, and also to the fact that a human benchmark [9] revealed that identifying African varieties in Broadcast News (our main source of data) is much harder than identifying accents of everyday's people on the street.

In this work, we take advantage of the existence of Portuguese variety-dependent systems to propose an automatic transcription system of media contents in different varieties of Portuguese based on automatic variety identification. The next section describes the main characteristics of the three variety-dependent ASR systems: AP, BP and EP. The variety identification system is described in section 3. It is composed by the fusion of 4 sub-systems (one acoustic and three phonotactic based) and is particularly designed for improved performance in close language/variety detection tasks. In section 4, cross-variety experiments show the importance of Portuguese accent as a speaker variability factor in speech recognition with a multi-variety corpus. Then, we evaluate the proposed multi-variety transcription system in contrast to an oracle system that uses the correct variety-dependent system for each test segment. Finally, this document finishes with the conclusions in Section 5.

2 Portuguese Variety-Dependent Transcription

2.1 The AUDIMUS Speech Recognizer

Our in house speech recognition engine [10] is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs). A block diagram is shown in Figure 1.

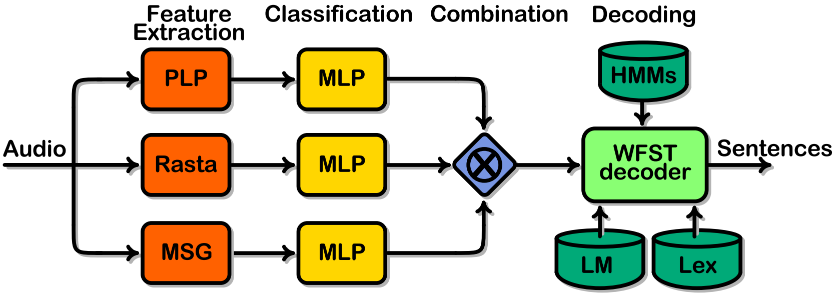

Fig. 1. Block diagram of AUDIMUS speech transcription system

Feature Extraction. The MLP/HMM acoustic model combines posterior phone probabilities generated by three phonetic classification branches. Different feature extraction and classification branches effectively perform a better modeling of the acoustic diversity, in terms of speakers and environments, present in data. The first branch extracts 26 PLP (Perceptual Linear Prediction) features, the second 26 Log-RASTA (log-RelAtive SpecTrAl) features and the 3rd uses 28 MSG (Modulation Spectrogram) coefficients for each audio frame.

MLP Classifiers and HMM Topology. Each MLP classifier incorporates local acoustic temporal context via an input window of several frames (between 7 and 15 frames) and is composed of two fully connected non-linear hidden layers and an output layer. Usually, the hidden layer size depends on the amount of training data available, while the number of softmax outputs of the output layer depends on the characteristic phonetic set of each language and the HMM topology. In the first versions of our recognizer single state context independent phoneme HMMs were trained, while in more recent versions a combination of multiple sate context independent units and intra-word context-dependent phone transition units are modeled [11].

Decoding Process. The decoder is based on the Weighted Finite-State Transducer (WFST) approach, where the search space is a large WFST that results from the integration of the HMM/MLP topology transducer, the lexicon transducer and the language model one [12]. This decoder uses a specialized WFST composition algorithm of the lexicon and language model components in a single step. Furthermore it supports lazy implementations, where only the fragment of the search space required in runtime is computed.

2.2 EP Transcription System

The EP transcription system was the first to be developed. Additional details can be found in [5].

EP Acoustic Model. The initial EP acoustic model was trained with 46 hours of manually annotated BN data collected from the public Portuguese TV. Currently, automatically collected and transcribed data is being reused to perform unsupervised training. The current iteration uses a total of 1000 hours of data mostly news shows from several EP TV channels. The EP MLPs are formed by 2 hidden layers with 2000 units each and have 500 softmax output units that correspond to 38 three state context independent phonemes of the EP language plus a single-state non-speech model (silence) and 385 phone transition units which were chosen to cover a significant part of all the transition units present in the training data.

EP Language Model. The Language Model (LM) is a statistically 4-gram model and results from the interpolation of three specific LMs. The first is a backoff 4-gram LM, trained on a 700M word corpus of newspaper texts, collected from the Web from 1991 to 2005. The second LM is a backoff 3-gram LM estimated on a 531k word corpus of broadcast news transcripts. The third model is a backoff 4-gram LM estimated on recent EP web newspapers texts, which are daily updated. The final interpolated LM was smoothed using Kneser-Ney modified discounting.

EP Vocabulary and Pronunciation Lexicon. The EP engine uses a 100k word vocabulary, which is also updated on a daily basis. The pronunciation lexicon is built automatically by classifying the words into "known" ones, for which the system retrieves a correct pronunciation from an in-house lexicon, and "unknown" ones. For the latter, a further split is made, which automatically detects spelled acronyms and foreign words. For "unknown" words which do not fit into these categories, the pronunciation is generated by our rule-based grapheme-to-phone (GtoP) conversion module [13]. For spelled acronyms, rule-based pronunciations are also generated. For foreign words, a further subdivision is made, in order to identify the ones that exist in the public domain lexicon provided by CMU[1], for which a nativized version is produced. For the words not included in the CMU lexicon, grapheme nativization rules are applied prior to using the GtoP module. The final multiple pronunciation EP lexicon includes 114k entries.

EP Results with BN Data. In one of our BN evaluation test sets (RTP07), which is composed by six one hour long news shows from 2007, our current EP BN transcription system achieves a word error rate (WER) performance of 18.4%.

[1] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

2.3 BP Transcription System

The BP transcription system is the result of porting some of the key modules of the EP recognizer to cover the BP variety. Particularly, it was necessary to train new acoustic models based on BP data, to build new language models, and to develop a new GtoP module. Details of this process can be found in [7].

BP Acoustic Model. The initial set of acoustic models for BP was trained with about 15 hours of manually transcribed BN data recorded from the Record channel transmitted by cable TV in Portugal. Due to the reduced amount of data available compared to EP, the size of the two nonlinear hidden layers of the MLPs was reduced to 600 units and only monophone units were modeled (39 phonemes + silence). Afterwards we increased the training data with more 33 hours of automatically transcribed material, which allowed increasing the size of the hidden layers to 1000 units and modeling multiple-state and phone transition units (up to a total of 320 softmax outputs).

BP Language Model. The BP language model is a 4-gram backoff model created by interpolating three individual LMs built from three different sources: the CETENFolha corpus which has around 24M words, a recent newspapers corpora automatically obtained from the Internet which amounts to 62M words and the manual transcriptions of the training set. The language models were smoothed using Kneser-Ney discounting and entropy pruning.

BP Vocabulary and Pronunciation Lexicon. BP recognizer also uses a 100K word vocabulary that includes all the words of the transcriptions of the training corpus, completed with the most frequent words of the newspaper corpus. The pronunciation lexicon is generated similarly as for the EP, but with a specific GtoP module for the BP variety with new rules. In this module, the grapheme set was augmented with the symbol ü, covering words written before the recent orthographic convention, and the phoneme set was augmented with three symbols: the two affricate symbols [tʃ] and [dʒ] and an additional SAMPA symbol to take into account the different realization of "r's" in coda position most commonly found in BP. The symbol for [ə] was not included, neither was [l].

BP Results with BN Data. The current version of the BP transcription system achieves a WER of 21.6% in our BP test set formed by 13 short news shows, which amount to almost 2 hours of speech.

2.4 AP Transcription System

The AP transcription system was the last to be developed. In contrast to the BP, the GtoP module has not been ported to AP and our efforts were restricted to the acoustic and the language models. More details can be found in [8].

AP Acoustic Model. In terms of acoustic model training two distinct approaches were followed. The first one consisted of training new acoustic models using only the small amount of AP training data available which is around 7.5 hours and using around 17 hours of automatically detected and transcribed AP data [8]. The second approach consisted of adapting EP acoustic models using only the manually transcribed AP data. In this last case, MLPs of two hidden layers with 1500 units modeling single-state phonemes were considered. In practice, this last approach resulted the most convenient.

AP Language Model. The language model for AP is a 4-gram model created by interpolating three individual language models. The first is the same 4-gram LM from the EP, the second LM was built from recent AP newspapers automatically obtained from the Internet with 1.6M words and the third LM was built from the manual AP transcriptions of the training set which amount to 86k words. The language models were smoothed using Kneser-Ney discounting and entropy pruning.

AP Vocabulary and Pronunciation Lexicon. The same 100K word vocabulary and pronunciation lexicon of the EP system is used for AP.

AP Results with BN Data. The AP test set consists of 3 short BN shows of among 25 and 30 minutes each one. In this test set, the AP system achieves a WER performance of 23.7%. It is worth noting that the AP BN test set is quite challenging due to the high correlation between noisy spontaneous speech conditions and AP accented speech found in BN shows obtained from the RTP Africa channel. The reason is that in these BN shows the anchor usually speaks with an EP accent, while AP accented speech is most often found in out of studio interviews and reports.

3 Portuguese Variety Identification

The Portuguese variety identification system in this work follows the principals of the one in [8]. It combines a state of the art acoustic sub-system based on Gaussian supervectors with three variations of a recently proposed phonotactic approach that makes use of "specialized" tokenizers, known as mono-phonetic Phone Recognition followed by Language Modeling (PRLM) [14] approach.

For each identification sub-system, target variety models are trained with the Portuguese variety training data-set described in [15], which consists of a total of 4141 BN speech segments of different lengths from the three target varieties. The complete variety identification system is calibrated and assessed with the variety development data-set that consists of 1484 segments from speakers that are not in the variety training set: 610 AP, 462 BP and 412 EP segments [15].

3.1 Mono-Phonetic PRLM Sub-Systems

PRLM systems make use of phonetic classifiers to tokenize speech data into phonetic sequences. Then, for each target language/variety a different phonotactic n-gram language model is trained using all the phonetic sequences extracted from the training data of that particular language/variety. During test, the phonetic sequence of a given speech signal is obtained and the likelihood of each target n-gram model is evaluated to obtain target language/variety scores.

Our PRLM approach focuses on the phonetic classifier, trying to build a system with a highly specialized tokenizer that incorporates the differences between language/variety pairs at this level. To better characterize these differences, we divide all occurring phones in our varieties into the following two groups [16]:

1. mono-phones: phones in one language/variety, that overlap little or not at all with those in another language/variety.
2. poly-phones: phones that are similar enough across the languages/varieties to be equated.

Determining the set of phones which best characterize a certain variety, given its neighboring varieties, is not straightforward. Linguistic knowledge about the varieties' phonetic and phonological characteristics is crucial, but often not available, not sufficiently detailed or controversial. We use a computational method instead to find variety dependent unique phones. Binary multi-layer perceptrons (MLPs) are trained to discriminate between the same pairs of aligned phone classes. In this work, we have used 3 phonotactic sub-systems, one per each Portuguese variety pair: AP-EP, AP-BP and BP-EP. More details about the determination of the mono-phonetic units and the training of the mono-phonetic classifiers can be found in [15].

3.2 Gaussian Supervector Based Sub-System

Combining Gaussian mixture models (GMM) with Support Vector Machines (SVM) as a discriminative classifier [17], the so-called Gaussian supervector (GSV) approach, is a well-known state-of-the-art technique in the Language Identification field. In this work, we have built a GSV system based on mean supervectors: Maximum-a-Posteriori (MAP) adaptation of Gaussian means. The extracted features are Shifted Delta Cepstra of PLP-RASTA features [18]. The universal background model (UBM) of 1024 mixtures was trained with all Portuguese variety data. In this implementation, we have used an alternative scoring approach [19]. In contrast to the conventional GSV, each language SVM model is *pushed back* to a *positive* and a *negative* variety-dependent GMM model, which are then used to calculate log-likelihood ratio scores. In certain situations, especially on short utterances, this approach has shown improved accuracy. In fact, this is the typical situation in BN data where long speech segments are rare.

3.3 Back-End Calibration and Identification Results

A linear logistic regression back-end for simultaneous fusion and calibration of the detection sub-systems has been developed using the FoCal Multiclass Toolkit[2]. Five-fold cross-calibration strategy was applied for back-end parameter estimation.

[2] http://niko.brummer.googlepages.com/focalmulticlass

The complete system generates three variety-dependent detection calibrated scores for every test segment. Since we are interested in variety identification rather than in verification, we select as the identified variety the one with the largest variety-dependent detection score. The miss probability and false alarm averaged results for the three varieties in the development set are 7.53% and 3.74% respectively, which corresponds to an average variety identification cost of 5.63%. Notice, however, that these might be optimistic performance indicators, since the same data set was used for both calibration and evaluation.

4 Multi-variety Transcription

4.1 Multi-variety Evaluation Corpus

The multi-variety evaluation corpus is formed by three subsets extracted from each one of the previously described BN speech recognition evaluation sets of the three Portuguese varieties. Concretely, for each variety we have randomly selected speech segments with a total approximate duration of around 45 minutes, totaling 131 minutes of useful speech. Table 1 summarizes the most relevant characteristics of the multi-variety corpus.

Table 1. Multi-variety corpus

Test Data	AP	BP	EP	$\sum$
duration [min.]	45.0	44.9	41.8	131.7
segments	282	505	360	1147
words	6948	7641	7148	21737
∅ dur./segm. [s]	9.6	5.3	6.9	6.9

4.2 Cross-Variety Speech Recognition Tests

Table 2 shows the WER performance results obtained by each variety-dependent transcription system, including results for each separate variety data subset and for the whole multi-variety corpus.

Table 2. Variety matched and cross-variety WER results

	AP	BP	EP	all
AP-ASR	**24.5**	49.0	22.4	32.4
BP-ASR	52.2	**22.1**	62.1	44.8
EP-ASR	27.2	57.0	**16.7**	34.2

Attending to the partial results, it is clear that the best performance is always obtained in matched variety conditions with a considerably robust performance in the three Portuguese varieties. In the ideal case of knowing the variety of each test segment (oracle system), the WER achieved in the overall multi-variety test set is 21.1%.

With respect to the cross-variety figures, it can be observed that AP and EP are much closer among them than the BP variety in terms of ASR performance. Thus, we can even observe in the AP-ASR row a better WER performance in the EP subset than in the AP subset. First, it is worth recalling that the AP ASR system makes use of acoustic models adapted from EP, language models that include EP information and that the same pronunciation lexicon is used. Second, as noted previously in Section 2.4, this figure also reveals that the AP subset is more challenging in terms of ASR than the EP one. With respect to the BP variety, although it is the "most distant" one in terms of cross-variety ASR results, these figures also seem to show a certain closer proximity to AP than to EP variety.

Regarding the overall results, the best variability-dependent transcription system is the AP one with a WER of 32.4%. However, the performance of the best individual system is far of the oracle system results (21.1% WER).

4.3 Multi-variety Recognition Based on Variety Identification

In our approach to the transcription of multi-variety Portuguese media contents, we first apply the variety identification system in a per segment basis in order to select the appropriate variety-dependent transcription system. Then, the selected transcriber is applied for that particular test segment.

Variety Identification. Table 3 shows the variety identification results obtained in the multi-variety test corpus with the system described in previous Section 3. In this case, the average variety miss and false alarm rates are 22.4% and 10.0% respectively, which corresponds to an average identification cost of 16.2%. These results represent a strong degradation with respect to the reference identification results reported in Section 3.3. This difference may be partially explained by the optimistic method used to calibrate and measure the performance of the variety ID system. Particularly, there is a very significant increase in the number of EP segments that are misclassified as AP, which results in a considerably boost of EP miss rate and AP false alarm rate. This fact seems to indicate the existence of a strong mismatch between the multi-variety EP sub-set and the characteristics of the EP data used for variety ID training and calibration. On the other hand, the average identification cost for the BP variety is of 6.8%, which is reasonably low. In spite of the performance drop with respect to the reference, these results are still quite encouraging since misclassification happens most often among varieties that are closer between them according to the cross-variety WER results reported in Table 2. Consequently, we expect a low impact in terms of multi-variety performance transcription.

ASR Results. Table 4 shows the results of the oracle system and of the proposed multi-variety system that makes use of automatic variety identification to select the variety-dependent recognizer. Attending to the individual variety subsets, an almost equivalent performance can be observed for AP and BP with respect to the oracle. In the case of AP, this occurs in spite of the fact that the

Table 3. Portuguese variety identification results in terms of false alarm rates *Pfa(Lt,Ln)*, where *Lt* is the target variety and *Ln* de actual variety, and miss probability rate *Pmiss(Lt)* and average false alarm rate *avg Pfa(Lt)* for each target variety

100 x *Pfa(Lt,Ln)*		*Lt*	
Ln	AP	BP	EP
AP	—	7.09	4.61
BP	5.74	—	0.20
EP	41.39	8.06	—
100 x *Pmiss(Lt)*	11.7	5.94	49.4
100 x *avg Pfa(Lt)*	20.58	7.63	1.8

miss rate for AP is 11.7%. It is likely that the segments erroneously classified as EP correspond to slight accented AP speech, which may be recognized equivalently or even better with the EP transcriber. The low miss rate for BP explains the low degradation of the multi-variety transcription system for this variety. On the other hand, the largest performance drop is observed in the EP subset and it is related to large misclassification of EP segments as AP segments. Anyway, the impact is not dramatic since, as already noted, the AP transcriber shares several components from the EP system. With respect to the overall result of the multi-variety transcriber, a performance of 22.7% WER is achieved, which is very close to the oracle system and much better than any of the individual variety-dependent recognizers of Table 2.

Table 4. WER results of the oracle and of the proposed multi-variety transcriber

	AP	BP	EP	all
oracle ASR	24.5	22.1	16.7	**21.1**
multi-variety ASR	24.5	22.6	21.0	**22.7**

5 Conclusions

In this work we have addressed the speaker accent variability issue for the transcription of media contents in broad varieties of the Portuguese language. First, we have demonstrated the strong impact of Portuguese varieties into speech recognition performance through a set of cross-variety experiments. Then, we have proposed a multi-variety transcription system based on the combination of variety identification and variety-dependent automatic transcribers. The proposed system showed excellent results when compared to the oracle system that uses true variety identity information. In the future we expect to improve the system, reducing the misclassification rate for the African/European Portuguese varieties pair and improving African Portuguese speech recognition. For this purpose, some possibilities include augmenting the amount of transcribed training data, namely for AP, and building a specific pronunciation lexicon or developing separate complete speech recognizers for each PALOP country variety.

Acknowledgments. This work was partially supported by FCT (INESC-ID multi-annual funding) through the PIDDAC Program funds, and also through the EU-funded project *EUTV Adaptive Media Channels*.

References

1. Huang, C., Chen, T., Li, S., Chang, E., Zhou, J.L.: Analysis of speaker variability. In: Proc. European Conference on Speech Communication and Technology, Denmark, vol. 2, pp. 1377–1380 (2001)
2. Huang, C., Chang, E., Chen, T.: Accent Issues in Large Vocabulary Continuous Speech Recognition. Microsoft Research China Technical Report, MSR-TR-2001-69 (2001)
3. Wang, Z., Schultz, T., Waibel, A.: Comparison of acoustic model adaptation techniques on non-native speech. In: Proc. ICASSP 2003, pp. 540–543 (2003)
4. Humphries, J.J., Woodland, P.C., Pearce, D.: Using accent-specific pronunciation modelling for robust speech recognition. In: Proc, Fourth International Conference on Spoken Language, ICSLP, vol. 4, pp. 2324–2327 (1996)
5. Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D.: Broadcast news subtitling system in Portuguese. In: Proc. ICASSP 2008, Las Vegas, USA (2008)
6. Lewis, M.P.: Ethnologue: Languages of the World, 16th edn., SIL International, (May 2009), `http://www.ethnologue.com/`
7. Abad, A., Trancoso, I., Neto, N., Viana, M.C.: Porting an European Portuguese broadcast news recognition system to Brazilian Portuguese. In: Proc. Interspeech 2009, Brighton, UK (2009)
8. Koller, O., Abad, A., Trancoso, I., Viana, C.: Exploiting variety-dependent phones in portuguese variety identification applied to broadcast news transcription. In: Proc. Interspeech 2010, Makuhari, Japan (2010)
9. Rouas, J., Trancoso, I., Viana, C., Abreu, M.: Language and variety verification on broadcast news for Portuguese. Speech Communnication 50(11-12), 965–979 (2008)
10. Meinedo, H., Abad, A., Pellegrini, T., Trancoso, I., Neto, J.: The L2F Broadcast News Speech Recognition System. In: Proc. Fala 2010, Vigo, Spain (2010)
11. Abad, A., Neto, J.: Incorporating acoustical modeling of phone transitions in an hybrid ANN/HMM speech recognizer. In: Proc. Interspeech 2008, Brisbane, Australia, pp. 2394–2397 (2008)
12. Caseiro, D., Trancoso, I.: A specialized on-the-fly algorithm for lexicon and language model composition. IEEE Transactions on Audio, Speech and Lang. Proc. 14(4) (2005)
13. Caseiro, D., Trancoso, I., Oliveira, L., Viana, C.: Grapheme-to-phone using finite state transducers. In: Proc. 2002 IEEE Workshop on Speech Synthesis, Santa Monica, CA, USA (2002)
14. Zissman, M.A.: Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. IEEE Transactions on Speech and Audio Processing 4(1) (1996)
15. Koller, O., Abad, A., Trancoso, I.: Exploiting variety-dependent phones in Portuguese variety identification. In: Odyssey 2010: The Speaker and Language Recognition Workshop (2010)

16. Berkling, K., Arai, T., Barnard, E.: Analysis of Phoneme-Based features for language identification. In: Proc. ICASSP, vol. 1, pp. 289–292 (1994)
17. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., Torres-Carrasquillo, P.A.: Support vector machines for speaker and language recognition. Computer Speech and Language 20(2-3), 210–229 (2006)
18. Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller Jr., J.R.: Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features. In: Proc. ICSLP 2002, Denver, Colorado, pp. 89–92 (2002)
19. Campbell, W.M.: A covariance kernel for svm language recognition. In: Proc. ICASSP 2008, pp. 4141–4144 (2008)