Language and Variety Verification on Broadcast News for Portuguese

Jean-Luc Rouas, Isabel Trancoso, Céu Viana, Mónica Abreu

Elsevier Editorial System(tm) for Speech Communication

Manuscript Draft

Manuscript Number:

Title: Language and Variety Verification on Broadcast News for Portuguese

Article Type: Special Issue: IBELAN

Section/Category: Special Issue Paper

Keywords: Language verification
Portuguese varieties

Corresponding Author: Dr. Isabel Trancoso,

Corresponding Author's Institution: INESC-ID

First Author: Jean-Luc Rouas , Doctor

Order of Authors: Jean-Luc Rouas , Doctor; Isabel M Trancoso, Doctor; Ceu Viana, Doctor; Monica Abreu

Abstract: This paper describes a language/accent verification system for Portuguese, that explores different type of properties: acoustic, phonotactic and prosodic. The two-stage system is designed to be used as a pre-processing module for the Portuguese Automatic Speech Recognition (ASR) system developed at INESC-ID. As the ASR system is applied everyday to transcribe the evening news from a Portuguese public TV channel, the presence of other languages (mainly English) and other varieties of Portuguese is very likely.
In the first stage, for each automatically detected speaker, the system verifies if the spoken language is Portuguese, as opposed to nine other languages -- English, Belgian Dutch, Croatian, Czech, Galician, Greek,Hungarian, Sloven and Slovak. The identified Portuguese speakers are then fed to the second stage which aims at identifying the Portuguese variety: European, Brazilian or African Portuguese from 5 countries. The identification results are then used either to mark the speech data as untranscribable or forward it to the European Portuguese ASR system, or a system tuned for other languages or varieties. The language verification system achieved an equal error rate for European Portuguese of 2.5%. In terms of variety identification, the overall rate of correct identification was 69.0\% if all 7 varieties are considered, and the best results were obtained for Brazilian Portuguese, also the variety that proved easiest to identify in perceptual experiments. If all African varieties are merged into a single broad class, the identification rate goes up to 94.7%. The fact that the prosodic system alone can achieve an identification rate of 77% is also worth investigating.

Title: Language and Variety Verification on Broadcast News for Portuguese

Authors:

Author 1: Jean-Luc Rouas, INESC-ID / INRETS
Address: INRETS Electronic, Waves and Signal Processing Research Laboratory for Transport, 20 Rue Élisée Reclus, 59650 Villeneuve d'Ascq, France
Email: rouas@inrets.fr

Author 2: Isabel Trancoso, INESC-ID / IST (Corresponding author)
Address: INESC-ID, Spoken Language Systems Laboratory (L2F)
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
Email: Isabel.Trancoso@inesc-id.pt

Author 3: Ceu Viana, CLUL
Centro de Linguistica da Universidade de Lisboa,
Av. Prof. Gama Pinto, 2, 1649-003 Lisboa - Portugal
Email: mcv@clul.ul.pt

Author 4: Monica Abreu, INESC-ID/IST
Address: INESC-ID, Spoken Language Systems Laboratory (L2F)
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
Email: monica7abreu@gmail.com

Abstract:

This paper describes a language/accent verification system for Portuguese, that explores different type of properties: acoustic, phonotactic and prosodic. The two-stage system is designed to be used as a pre-processing module for the Portuguese Automatic Speech Recognition (ASR) system developed at INESC-ID. As the ASR system is applied everyday to transcribe the evening news from a Portuguese public TV channel, the presence of other languages (mainly English) and other varieties of Portuguese is very likely. In the first stage, for each automatically detected speaker, the system verifies if the spoken language is Portuguese, as opposed to nine other languages -- English, Belgian Dutch, Croatian, Czech, Galician, Greek, Hungarian, Sloven and Slovak. The identified Portuguese speakers are then fed to the second stage which aims at identifying the Portuguese variety: European, Brazilian or African Portuguese from 5 countries. The identification results are then used either to mark the speech data as untranscribable or forward it to the European Portuguese ASR system, or a system tuned for other languages or varieties. The language verification system achieved an equal error rate for European Portuguese of 2.5%. In terms of variety identification, the overall rate of correct identification was 69.0% if all 7 varieties are considered, and the best results were obtained for Brazilian Portuguese, also the variety that proved easiest to identify in perceptual experiments. If all African varieties are merged into a single broad class, the identification rate goes up to 94.7%. The fact that the prosodic system alone can achieve an identification rate of 77% is also worth investigating.

Language and Variety Verification on Broadcast News for Portuguese

Jean-Luc Rouas[1,2], Isabel Trancoso[1], Céu Viana[3], Mónica Abreu[1]

[1] *INESC-ID, Spoken Language Systems Laboratory (L2F), R. Alves Redol , 9, 1000-029 Lisboa, Portugal*

[2] *INRETS Electronic, Waves and Signal Processing Research Laboratory for Transport, 20 Rue Élisée Reclus, 59650 Villeneuve d'Ascq, France*

[3] *Centro de Linguística da Universidade de Lisboa, Av. Prof. Gama Pinto, 2, 1649-003 Lisboa - Portugal*

Abstract

This paper describes a language/accent verification system for Portuguese, that explores different type of properties: acoustic, phonotactic and prosodic. The two-stage system is designed to be used as a pre-processing module for the Portuguese Automatic Speech Recognition (ASR) system developed at INESC-ID. As the ASR system is applied everyday to transcribe the evening news from a Portuguese public TV channel, the presence of other languages (mainly English) and other varieties of Portuguese is very likely. In the first stage, for each automatically detected speaker, the system verifies if the spoken language is Portuguese, as opposed to nine other languages – English, Belgian Dutch, Croatian, Czech, Galician, Greek, Hungarian, Sloven and Slovak. The identified Portuguese speakers are then fed to the second stage which aims at identifying the Portuguese variety: European, Brazilian or African Portuguese from 5 countries. The identification results are then used either to mark the speech data as untranscribable or forward it to the European Portuguese ASR system, or a system tuned for other languages or varieties. The language verification system achieved an equal error rate for European Portuguese of 2.5%. In terms of variety identification, the overall rate of correct identification was 69.0% if all 7 varieties are considered, and the best results were obtained for Brazilian Portuguese, also the variety that proved easiest to identify in perceptual experiments. If all African varieties are merged into a single broad class, the identification rate goes up to 94.7%. The fact that the prosodic system alone can achieve an identification rate of 77% is also worth investigating.

Key words: Language verification, Portuguese varieties.

Email address: `rouas@inrets.fr`, `Isabel.Trancoso@inesc-id.pt`, `mcv@clul.ul.pt`, `monica7abreu@gmail.com` (Jean-Luc Rouas[1,2], Isabel Trancoso[1], Céu Viana[3], Mónica Abreu[1]).

The Spoken Language Systems Lab (L2F) of INESC-ID has been actively working on Automatic Speech Recognition (ASR) for many years. The Portuguese ASR system is currently applied to the transcription of the broadcast news extracted from a public national channel, the "Telejornal" on RTP1. The system is working on a daily basis and results of the transcription of the last broadcasted evening news are available at `http://www.l2f.inesc-id.pt/wiki/index.php/Demos`.

However, one of the problems encountered by the ASR system is the presence of different languages: many interviews are subtitled in Portuguese, while the audio remains in the original language. This generates a long stream of errors which can have a very negative impact on any modules that follow the recognition module (search, indexation, summarization, etc.). Therefore the system needs to know if the spoken language is really Portuguese or another language. Furthermore, if several ASR systems are available for the most frequent other languages (like English), this also allows the selection of the most appropriate ASR system. Moreover, in case the Portuguese language is identified, we also have to determine which variety of Portuguese is actually spoken, as there may be great variations in pronunciation.

This paper is organized as follows: We start by recalling the cues commonly used for language and accent characterization (section 1). Then we make a short review of state-of-the-art systems for language and accent identification (section 2). Section 3 is dedicated to a discussion about the differences between Portuguese varieties. The design of the language verification system is detailed in section 4. The corpora used for the experiments on language and variety identification are described in section 5. Experiments on language verification are discussed in section 7, focusing mainly on the performances of the Portuguese language verifier. In section 8, we study how the language verification system reacts when provided with samples of different varieties of Portuguese. Finally, the performance of the variety verification system is discussed in section 9.

1 Introduction

The aim of automatic language identification (LID) is to find which language is spoken in an utterance spoken by an unknown speaker. Several cues can be used for this purpose, based on linguistic and perceptual studies on the differences among languages.

In [46], four kinds of cues are described:

- Phonology: The phoneme sets used in different languages differ, even though many language share a common subset. Phonotactics, i.e. the rules governing the sequences of phonemes are also different.
- Morphology: The words roots and lexicons differ from one language to another.

Each language has its own vocabulary.
- Syntax: The way to construct sentences is different among languages.
- Prosody: Rhythm and intonation patterns are different.

Phonological properties are used in the most common language identification systems:

- The *Acoustic Language Identification Systems* use the differences in acoustic realizations of phonemes.
- The *PRLM (Phone recognition followed by language modeling) Systems or PPRLM (Parallel-PRLM) Systems* characterize each language by its most frequent sequences of phones.

Morphological and syntactical cues are hard to deal with if we do not have the transcription output of a speech recognition system, which is a language-specific task.

Prosody is also hard to model, mostly because of the suprasegmental nature of the prosodic features. That is why prosody has seldom been used for high performance language identification. More recently, however, prosodic features are beginning to be integrated in many systems, conjointly with acoustic or phonotactics, in order to take into account all the available information [44].

2 Actual performances for Language verification systems

Despite the growing interest on language identification that was observed during the 1980s, this area has not been much considered in the following decade. Nowadays, there is a regain of interest for language identification systems, probably motivated by their potential application in surveillance. This interest also led to significant improvement in performances as shown by the more recent NIST evaluations. The task addressed by these evaluations are however different from "classic" language identification experiment.

Traditionally, the language identification systems were asked to identify a language within a finite set of languages. Since the 1996 NIST Language Recognition Evaluation, the task has moved to language *verification*, which is similar to speaker verification: the aim is to evaluate if the speech excerpts belong to a target language, or not.

As the NIST evaluations (the next one will take place in September 2007) are a very good ground for estimating language verification system performances, some of the best performing systems of the latest 2005 evaluation are briefly described below (see `www.nist.gov/speech/tests/lang/` for a complete list of participating organizations).

The languages used in the NIST 2005 Evaluation were the following: English (American & Indian), Hindi, Japanese, Korean, Mandarin (Mainland & Taiwan), Spanish (Mexican) and Tamil. For the "primary condition" (i.e. mandatory experiments), only the American English and Mainland Mandarin varieties were considered. Thus, the corpus used for evaluation is composed of 7 languages.

The evaluation of the system is achieved using Equal Error Rate (or EER), which means balanced errors between false alarms and missed detections. Usually, a detection error trade-off curve (or DET-curve) is also provided as a characteristic of the performances of the tested system.

Of the several language verification systems used in the NIST 2005 Evaluation campaign, the best performing ones use either acoustic or phonotactic (P-PRLM) approaches or a fusion of both.

For example, the Brno university system (described in [24]) uses a GMM-based acoustic system combined with a Neural Network-based PPRLM system. Combining the scores of these approaches with a weighted addition of the log-likelihoods give an overall equal error rate of 5.0% on 30 seconds excerpts.

The system submitted by the Georgia Institute of Technology and the Infocomm Institute uses a fusion of two approaches [22]. The first one is a classical PPRLM. The second approaches uses a "Bag of Sound" (BOS) recognizer, which can be also called "universal phone recognizer". This BOS recognizer is trained to recognize 258 phonemes from 6 languages (English, Mandarin, Japanese, Hindi, Spanish and German). Then, SVM classifiers are used to make pairwise decisions. The scores obtained from both approaches are concatenated to form a score vector which is fed to the back-end system. Two approaches are used to provide scores for each of the 16 target languages: Artificial Neural Networks or Linear Discriminant Functions. The results obtained by each of these classifiers are then merged. The performance obtained with this system is 12% EER on the NIST 2005 30 seconds utterance data set.

The Lincoln Laboratory of the Massachusetts Institute of Technology has presented a fusion of several systems [6]. The systems were: GMM-SDC (Gaussian Mixture Models with Shifted Delta Cepstra Features), SVM-SDC (Support Vector Machine with Shifted Delta Cepstra Features), PPRLM (Parallel Phone Recognition followed by n-gram Language Models classifiers), PPRLM-lattice (Parallel Phone Recognition followed by n-gram Language Models classifiers using Phone Lattices [15]), PPRSVM-lattice (Parallel Phone Recognition followed by Support Vector Machine classifiers using Phone Lattices), and PPRBT (Parallel Phone Recognition followed by Binary Tree Language Models (developed at IBM)). The fusion is achieved by modeling the concatenated output scores of each of these systems by Gaussian Mixture Models. The performances reached by this system is 4.2% of

Equal Error Rate on 30 seconds test utterances.

All these systems show the performance that is achieved nowadays on the language verification task. While being the best performing systems, PPRLM are also the most complex ones (both in terms of design and computational time). In fact, building a powerful PPRLM system almost requires the implementation of speech recognizers for several languages. The acoustic modeling systems have been thoroughly investigated during last years, taking benefits from speaker verification researches, and are now almost competitive with PPRLM systems. It is however noticeable that none of these systems use prosodic features.

Dialect identification is a somewhat harder topic that language identification and has not been for the moment as much investigated [23] [4] [14] [36] [43] [45] [17] [16] although one can find a growing number of references on a related problem - foreign accent identification [42]. Many approaches use language identification systems applied to native dialect identification.

For example, in [39] a GMM-SDC based system is applied to Spanish dialects identification, considering only two dialects (Cuban and Peruvian) with the "Miami" corpus. On this data, the system generates an error rate over 30%. This experiment has also been carried on the dialects present in the CallFriend corpus, using 30 seconds utterances: American English (North vs. South), Chinese (Mandarin vs. Taiwan) and Spanish (Caribbean vs. Non-Caribbean). The error rate were respectively: 15% for American English, 11.5% for Chinese and 13.7% for Spanish.

In [8], another GMM-based system is applied to Chinese dialect identification. The accents present is this corpus come from 4 regions: Beijing, Shanghai, Guangdong and Taiwan. The data used come from the "Multi accent Mandarin corpus", consisting in 1440 speakers for approximately 16 hours. 60 speakers were used for testing for each dialect. The results were between 12% and 15% errors (for female and male speakers respectively) for utterance of approximately 20 seconds.

The experiments reported in [40] concern also the identification of Chinese dialects. Here the considered dialect are Mandarin, Holo and Hakka (all spoken in Taiwan). The corpus used in these experiments is quite small, with a total of 8 speakers reading 30 paragraphs, generating sentence about 15 seconds long. The same speakers are used for training and testing, and each speaker read each text 3 times, once in each of the dialects. Using MFCC and pitch features and a Gaussian mixture bigram model, the system achieves a performance of 94% of correct identifications. These experiments however show the importance of considering prosodic information, as using only the pitch-based features, the identification rate is 57%.

The prosodic system developed in the PhD thesis of the first author[32] was successfully tested on read speech from 7 languages (English, French, German , Italian, Spanish from the original Multext corpus [7] and Mandarin Chinese [19] and Japanese [18]) – around 70% of correct identifications, experiments described in

[33] – and semi-spontaneous Arabic dialects (on the Araber database [35]). On this database, the task was to discriminate between geographical areas linked to the dialects in 3 zones: Maghreb, Middle-East and Intermediate (Egypt and Tunisia) and the area identification rate was 98%.

Unfortunately, we do not have the same experience on Portuguese dialect identification. In the next section, we will describe the main differences between the Portuguese varieties and discuss how we can take them into account in our system.

3 Main differences between the varieties of Portuguese

This section summarizes the main differences between some of the varieties spoken in CPLP countries (Community of Portuguese-speaking Countries). Portuguese is a language that is spoken by more than 170 million people in virtually all continents, ranking it very high among the most spoken languages in the world. The current work does not cover all of them, being restricted to the varieties to which we could have easy access in term of Broadcast News (BN) recordings [1] :

- European Portuguese (henceforth denoted as EP), the variety spoken in Portugal, for which the available speech recognition system has been trained.
- Brazilian Portuguese (henceforth denoted as BP), the variety spoken in Brazil, with the largest number of speakers.
- African Portuguese (henceforth denoted as AP), the generic name that covers all the varieties spoken in African countries that have Portuguese as official language (PALOP countries): Angola (AN), Cape Verde (CV), Guinea-Bissau (GB), Mozambique (MO) and São Tomé and Príncipe (ST).

Whereas there are already quite a few reports on the differences between EP and BP, the differences between these varieties and AP are much less studied. Many of the comments made in this paper concerning AP will hence be made based on the study of the corpus described in Section 5. Unfortunately, Broadcast News is not the type of controlled conditions corpus that should ideally be used for this purpose. Speakers from African countries often have Portuguese as second language (namely in rural areas), and we cannot guess the native language in such multilingual environments. Their education degree is also very variable, as is the contact they may have with other varieties of Portuguese. Hence our comments on AP are mostly preliminary and need further corroboration with more controlled corpora.

[1] Speakers from Timor were unfortunately very scarce in BN transmitted in Portugal

3.1 Orthographic and syntactic differences

The current orthographic convention allows for minor differences, representing some phonetic and phonological specificities: the optional suppression of unpronounced consonants in BP (e.g. *acção* / *ação*, *excepto* / *exceto*), the optional use of hyphenation, and differences in diacritics (e.g. *tranquilo* / *tranqüilo*, accounting for the fact that *u* is pronounced as /w/, instead of deleted as in the general case involving *qui* or *que* sequences; *Jerónimos* / *Jerônimos*, accounting for the different vowel quality).

Besides these differences, there are also significant ones concerning the use of prepositions, the position of clitics and the alternative use of infinitive/gerundive verb forms (e.g. *estava sempre a meter-se em sarilhos* vs.*estava sempre se metendo em sarilhos* - was always getting into trouble).

African countries that have Portuguese as official language follow the same orthographic conventions as for EP. Although the written form is very similar in AP and EP, in spontaneous speech in AP one can find very frequent instances of lack of number agreement (e.g. *os joelho* instead of *os joelhos* 'the knees'). The causes for this phenomenon, which can also be found in BP, are controversial. Some authors relate it to the influence of Bantu languages, where the plural form does not need to be marked in both the determinant and the noun, as in the example above.

3.2 Phonetic and phonological differences

There is common agreement that one of the most striking differences between Brazilian and European varieties concerns vowel reduction, which is much more extreme in EP than in BP [25], [3]. EP unstressed high vowels are often deleted and rather long consonant clusters may surface within as well as and across word boundaries, which are not allowed in BP (e.g. *se desprezarmos* [sdʃprzˈarmuʃ] 'if we ignore'). As empty nuclei are also obligatorily filled in BP, most two-obstruent sequences are broken by an epenthetic vowel (e.g. *psicologia* [pisikoloʒˈiɐ] 'psychology', *afta* [ˈafitɐ] 'aphtha' in BP vs [psikluʒˈiɐ], [ˈaftɐ] in EP). Loanwords can be treated rather differently, as well (eg. [iʒnˈɔbi] in BP vs [snˈɔb] in EP). Although both varieties distinguish between seven vowels in stressed position (/i e ɛ a ɔ o u/), they do not have the same reduction patterns, and quality changes are not sensitive to the same constraints.

The number of contrasting vowels is context dependent in BP: in pre-tonic position, /e/-/ɛ/ and /o/-/ɔ/ contrasts are neutralised and the seven-vowel system reduces to the five-vowel system /i e a o u/, whereas in post-tonic position, it reduces to the three vowel system /i ɐ u/, as /i e ɛ/ and /u o ɔ/ merge to [i] and [u] respectively, and /a/ is raised to [ɐ]. EP does not show this type of variation, as its four-vowel system (/i ə ɐ u/) holds for both positions.

In BP, unstressed vowels must also agree in height with the word stressed vowel (e.g. *preferência* (preference) [preferˈẽsjɐ] - *preferível (preferable)* [prifirˈivew]. Vowel height harmony in BP has been extensively studied, as it constitutes an important factor for the differentiation of BP dialects [5], [21]. According to these authors, it is a variable rule, which mainly affects the vowel immediately adjacent to the stressed one, and whose application depends on a multiplicity of factors (such as the presence/absence of a front vowel in the stressed syllable, presence/absence of a morphological boundary, speaker's age, etc.). Vowel lowering is typical of northern dialects and is practically nonexistent in Rio and S. Paulo. As for the raising of mid vowels, the authors found that harmony is respected in 32% and 29% only, for [e] and [o] respectively.

Although stressed vowels are rather similar in both varieties, there are some small differences worth mentioning. In EP, an additional vowel ([ɐ]) may also appear in this context, as in some dialects including the Lisbon one:

(1) [ɐ]/[a] distinguish between the 1st person plural of verbal present and past tense forms, respectively (e.g. *pensamos* (we think) [pẽsˈɐmuʃ] / *pensámos* (we thought) [pẽsˈamuʃ]);
(2) low vowels are raised before heterosyllabic nasal consonants (e.g. *cara* (face) [kˈarɐ] / *cana* (cane) [kˈɐnɐ]);
(3) front vowels centralise before palatal consonants and glides (e.g. *desenho* (drawing) [dzˈɐɟu], *telha* (tile) [tˈɐʎɐ], *lei* (law) [lˈɐj]).

In BP the two forms in (1) are homophones (*pensamos* [pẽsɐmus] or [pẽsɐmuʃ]), and provided the orthography is correct the desired pronunciation is generated. In fact, although since [20], it has often been pointed out that nasalization is much stronger in BP than in EP, it has also been shown [1] that it is not a categorical rule: full nasalization is favored in stress position (> 90% of the cases) but several factors, such as the presence of empty onsets or morphological boundaries, may inhibit nasal spreading in other contexts. On the other hand, some EP speakers may also strongly nasalize vowels in stressed position.

As for the main differences concerning the consonantal system, they are well known. In EP, coronal plosives are realized as [t] and [d], whereas in BP they are realized as [tʃ] and [dʒ], respectively, before /i/.

Coda consonants in BP may considerably differ from EP ones in the same context. In fact, the realization of the so-called strong and weak "r's" varies considerably across the country, namely in coda position. In coda position, "l" is realized as [ɫ] in EP, and as a labio-velar offglide, in BP. Due to this fact, a larger diphthong list can be found in BP.

In BP, diphthongs may also emerge from yodisation of some vowels followed by

/ʃ/, as in *arroz* (rice) [aɾˈojʃ].

Having summarized the main differences between EP and BP, which are fairly well studied, let us know address the much more unexplored comparison with AP varieties.

The multilingual background of many AP speakers may be the cause for the very large variability in the reduction patterns of AP varieties, both inter and intra-speaker. On one hand, one can find instances of vowel epenthesis in order to break consonant clusters and respect the CV syllable pattern. On the other hand, one can also find a generalization of EP reduction rules that, together with the influence of complex consonants of some native languages, may lead to patterns of vowel reduction even more extreme than those found in EP.

Similarly to BP, in AP consonant clusters formed across word boundaries may be solved in different ways: either by insertion of a paragogic vowel or by deletion of the coda consonant in the last syllable of the first word. In our data, the most common epenthetical vowel is /ə/, rather than /i/ as in BP. The latter occurs mostly in the last position of verbal forms ending in consonant. Contrarily to what is generally thought, there is no evidence that this vowel may be a copy of the following syllable nucleus. It is possible that, for other varieties such as observed for MO [2], the process is very frequent for borrowings of Portuguese words by native languages but not to dissolve clusters in Portuguese words.

Contrarily to BP, in AP vowels are often deleted between nasals and obstruent consonants and pre-nasalized onsets often occur (e.g. *amizade* 'friendship', pronounced as [ɐmzˈadə] instead of [ɐmizˈadə] as in EP). Deletion of high vowels or entire rhymes may also occur for AP within as well as across word boundaries, not only when the resulting sequences are similar to well-formed affricates, but also for other combinations of coronal fricatives with other obstruent consonants (e.g. *psicólogo* 'psychologist' often pronounced as [pskˈɔlugu] in AP compared to [psikˈɔlugu] in EP).

Concerning vowel reduction in pre-tonic position, a significant inter and intra-speaker variability is found in AP. Either there is vowel raising and centralization (sometimes more extreme than for EP) or there is a mixed behavior as some vowels are raised and others are not. This is often the case when a non-raised pre-tonic vowel would be produced with the same quality as the following stressed one.

Another very frequent phenomenon in AP is the neutralization of the /e/-/ɛ/ and /o/-/ɔ/ contrasts, but here, again, we have observed an enormous variability in all varieties.

The fact that some of the speakers do not contrast /e/-/ɛ/ and /o/-/ɔ/ in stressed

[2] F. Vicente, personal communication

position and realize both vowels in each pair with an intermediate quality suggests that a contrast may not occur in their native languages. Apparently, there seems to be a generalization of an EP-metaphony rule according to which /ɛ/ and /ɔ/ are realized as /e/ and /o/, respectively, in penultimate stressed open syllables, when the following syllable has a high rounded vowel (e.g. *mesa* 'table', pronounced as [mˈɛzɐ] in AP and as [mˈezɐ] in EP; e.g. *preso* 'arrested', pronounced as [prˈezu] in AP and EP]).

For all varieties but most noticeably in CV, unstressed /a/ is generally pronounced as /ɐ/, even in closed syllables in which vowel reduction is blocked in EP (e.g. *principalmente* 'mainly', pronounced as [prĩspɐlmˈẽtə] in AP and as [prĩsipalm̃ˈẽtə] in EP). Also, contrarily to EP, the fusion of two unstressed /a/ results in a single central vowel [ɐ] instead of in a low one ([a]).

Falling diphthongs tend to monothonguize, in particular nasal ones, and what should be rising diphthongs in EP tend to be pronounced in hiatus. In the latter case, instead of the glide, a lowered vowel may be found (e.g. *habituados*, 'used', pronounced as [ɐbitoˈaduʃ] in AP and as [ɐbitwˈaduʃ] in EP).

Coronal consonants are often apico-alveolar in all varieties. This is most noticeable for liquids. Some speakers do not produce a trill, neither in initial nor in intervocalic position.

3.2.1 Prosodic differences

The literature on the rhythm of Portuguese shows that there are controversial issues. In [29], for instance, EP is classified as stress-timed and BP as having mixed patterns of the syllable and stress-timed type. In [12], on the other hand, EP is characterized as having both stress-timing and syllable-timing properties, and BP as showing both syllable- and mora-timing properties. In a later paper [11], the same authors claim that EP and BP can be discriminated when the intonation pattern is preserved and all segmental information has been filtered out, and discuss the fact that intonation may be one of the important factors that lead to rhythmic distinctions, a topic that they view as worth pursuing.

Whereas comparative studies of BP and EP prosody can already be found (see also [10]), as far we know, such studies are inexistent for African varieties. However, we strongly believe that they will play a crucial role in distinguishing between themselves. In fact, the observation of our Broadcast News corpus allowed us to detect major differences between AP, BP and EP varieties from the segmental point of view, but these differences were more or less shared by all the AP varieties, with the above mentioned strong inter and intra-speaker variability. Subjects with some familiarity with the different African varieties are able to make a fair discrimination among them based on prosodic cues. The present work is a step towards studying these differences.

4 Language identification system

After the necessarily brief review of the most recent LID approaches, we have retained the following options:

- The PPRLM systems seem to achieve the best results, so it is relevant to implement one. The main difficulty is that PPRLM systems need several high-performance phone recognizers. As there is already a phone recognizer available for the Portuguese language, and as our system is mainly targeted at testing the presence of the Portuguese language in BN, we have decided to design a simple PRLM system using the Portuguese phone recognizer.
- The acoustic systems are an interesting compromise between complexity and performance. We have implemented a simple acoustic system using MFCC coefficients and Gaussian Mixture Models.
- As hypothesized by linguistic studies, prosody may also be a relevant cue to differentiate Portuguese varieties. Thus, taking in account our previous knowledge on prosody modeling and dialect identification [34,35], we have decided to implement a prosodic system, conjointly with the PRLM and acoustic system.

The system is thus a fusion of 3 subsystems: Acoustic (section 4.2) , Phonotactic or PRLM (section 4.2), and Prosodic (section 4.4). These 3 subsystems use a common pre-processing module as represented in the Figure 1. The pre-processing module will be briefly reviewed in the following section.

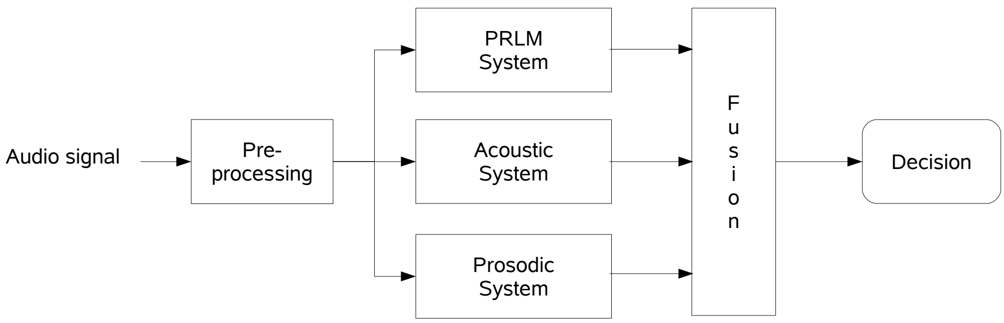

Figure 1. Overview of the language identification system.

For the time being, the fusion method is only a simple weighted addition. Hence, it will not be described in detail and is only mentioned to give an idea of the performances that could be achieved using the three subsystems together in the section on experimental results (section 7).

4.1 Pre-processing

The language identification system is designed to be integrated in the speech recognition system. Therefore, it is relevant to take advantage of the pre-processing module also used in the speech recognition system. Two of the blocks of this pre-processing module are of particular interest for language identification: the speech/non speech detection, as we do not want to treat non-speech parts, and the speaker clustering. The latter is used to take the language identification decision, as we make the hypothesis that a unique speaker speaks only one language.

The pre-processing system, developed by H. Meinedo (see [27] and [28]), is composed of five modules (Figure 2): three modules for classification (Speech/Non Speech, Gender and Background), one for speaker clustering and one for acoustic change detection. All the modules are model-based, that is to say they use algorithms trained using *a priori* information. These models are composed of Artificial Neural Networks (ANNs) of the type feed-forward fully connected Multi-Layer Perceptron (MLP), and were trained with the back-propagation algorithm.

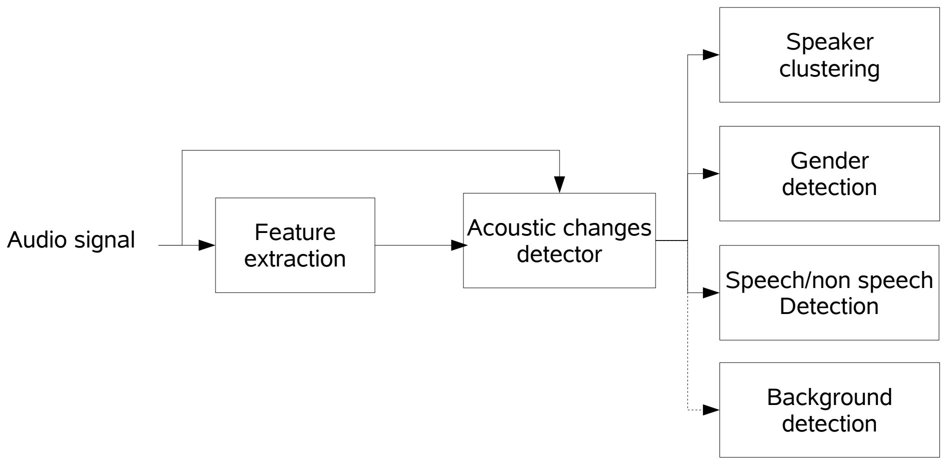

Figure 2. Overview of the pre-processing module.

The acoustic change detection module is used to divide the input audio stream into acoustically homogeneous segments. This is achieved using a hybrid two-stage algorithm, combining energy, metric and model-based techniques.

The classification modules are mainly used for helping the speaker clustering system: the Speech/Non speech classifier detects speech segments; the gender classifier is used to reduce the search space for speaker clustering; the background classifier aims at the identification of background conditions - clean, noise, music, etc.

After the acoustic change detector signals a new acoustic segment that is classified as a speech segment belonging to a male or a female speaker, the speaker clustering compares this segment to all the clusters found so far. The new segment is eventually merged with the cluster for which the lowest distance was calculated, below a predefined threshold. The distance computation is based on an MLP classifier trained to estimate the probability of the acoustic features extracted for a given segment to belong to a particular cluster. As the speaker clustering classification makes use of the gender information, two classifiers are built, one for each gender.

Two of the modules of this pre-processing stage are specially interesting for the language identification stage: the speech/non speech detection, as we do not want to treat non-speech parts, and the speaker clustering, as we assume that each speaker speaks a single language and make the language verification decision on a speaker by speaker basis.

4.2 Acoustic system

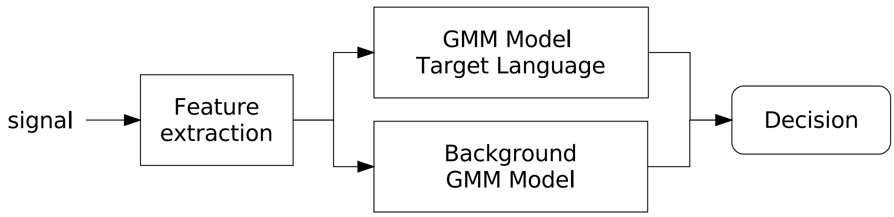

Figure 3. Generic acoustic language verification system.

A generic acoustic language identification system is displayed on Figure 3. The system works in two phases: a learning procedure to create the models, and testing procedure. The acoustic features extracted from the audio signal can be of different kinds (for example Mel Frequency Cepstral Coefficients, Perceptual Linear Prediction coefficients, or Shifted Delta Cepstra). In our case, the feature vector is composed of 12 MFCC plus delta, resulting in a 24-dimensional vector.

The models can also be of different nature (Gaussian Mixture Models, Artificial Neural Networks (ANN), Support Vector Machines, ...), using different learning algorithms (Vector Quantization (VQ) followed by Expectation Maximization (EM) for the GMM, Back-Propagation for ANN, ...). The decision can also be taken in different ways (Maximum of Likelihood, distance measure, ...) Here, the models used are Gaussian Mixture Models, learned with the classic VQ and EM algorithms.

The verification test is made by comparing the likelihood of the test excerpt to the target-language model and to the *background* model. This background model is learned using data from all available languages.

Hence, the system output consist in a decision (true or false) if the language spoken in the test excerpt is the target language, and a confidence score (the ratio of the likelihoods from the target language model and the background model).

4.3 PRLM system

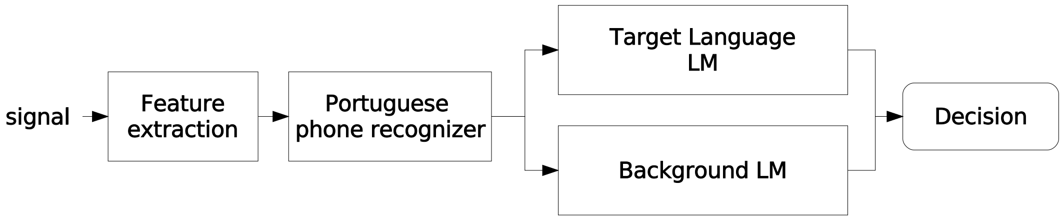

Figure 4. PRLM System overview.

As explained above, the PRLM system is based on a single Portuguese phone-recognizer. A synoptic of the system is given in the Figure 4.

The phone recognizer is part of the AUDIMUS system [26]. AUDIMUS is a hybrid system that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification abilities of Multi-Layer Perceptrons.

The phonetic decoding in the AUDIMUS system is based on MLP models. It combines phone probabilities computed from several MLPs using different feature sets: PLP (12^{th} order plus delta), log-RASTA (12^{th} order plus delta), and Modulation Spectrogram (MSG - 28 coefficients). These MLP classifiers incorporate context with an input comprising 9 frames (4 frames left and 4 frames right from the actual frame). The networks have a hidden layer of more than 1000 units and 40 output units corresponding to the context-independent phone classes (the 38 Portuguese phone set plus silence and breath noise). The outputs of all 3 MLP classifiers are then merged using an average in the log-likelihood domain.

This phonetic decoding is applied to all the languages in the training database, resulting in Portuguese-phones sequences which are then modeled for each language by n-grams, using the SRI-LM toolkit [38].

The verification test is made by comparing the likelihood of the test excerpt to the target-language model and to the *background* model. N-gram models are learned for the target language and the background model is learned using data from all languages. During the test phase, the identified language is defined according to the n-gram model providing the maximum of likelihood.

Like the acoustic system, the PRLM system output consist in a decision (true or false) if the language spoken in the test excerpt is the target language, and a confidence score.

4.4 Prosodic system

The prosodic system is the same as used in [34]. It is based on two different aspects: the definition of relevant units (pseudo-syllables) and the separate processing of the variations of macro- and micro-prosodic components (long- and short-term models). A synoptic of the system is displayed on Figure 5.

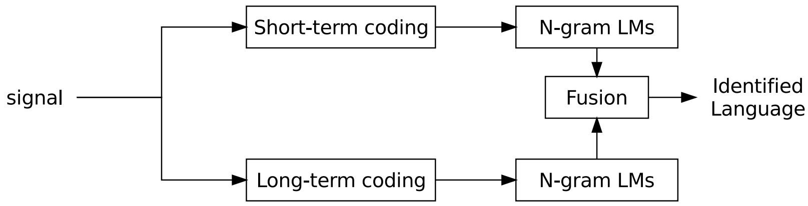

Figure 5. Prosodic system overview.

4.4.1 Segmentation, Vowel detection and Pseudo-syllables

The pseudo-syllable unit is defined as a cluster of consonants ending with a vowel, corresponding to the most frequent syllable structure in the world [9].

Three baseline procedures lead to relevant consonant, vocalic and silence segment boundaries:

- Automatic speech segmentation leading to quasi-stationary segments: This segmentation is issued from the "Forward-Backward Divergence" (DFB) algorithm [2], which is based on a statistical study of the signal in the temporal framework. The segmentation achieves an infra-phonemic segmentation where segments correspond to steady or transient parts of phonemes.
- Vocal activity detection: The vocal activity detection is based on a first order statistic analysis of the temporal signal [31]. The activity detection algorithm detects the less intense segment of the excerpt (in terms of energy) and the other segments are classified as Silence or Activity according to an adaptive threshold.
- Vowel localization: The vowel location algorithm is based on a spectral analysis (see [31] for more details). The fact that neither labeled data nor supervised learning are necessary constitutes the main advantage of this algorithm. The fact that no learning phase is necessary allows the algorithm to be used on different languages, even if no hand-labeled data is available. However, the consequence is that the algorithm is not optimized for any language even if it behaves correctly when compared to other systems [30].

This front-end processing results in a segmentation into vocalic, consonantal and silence segments. Labels "V", "C", or "#" are used to qualify each segment.

Once consonantal, vocalic and silence segments are identified, we can process the pseudo-syllable gathering: all the consonantal segments are merged until the next vocalic segment, which ends the pseudo-syllable. Figure 6 shows the automatic segmentation and labeling results and the identified pseudo-syllables.

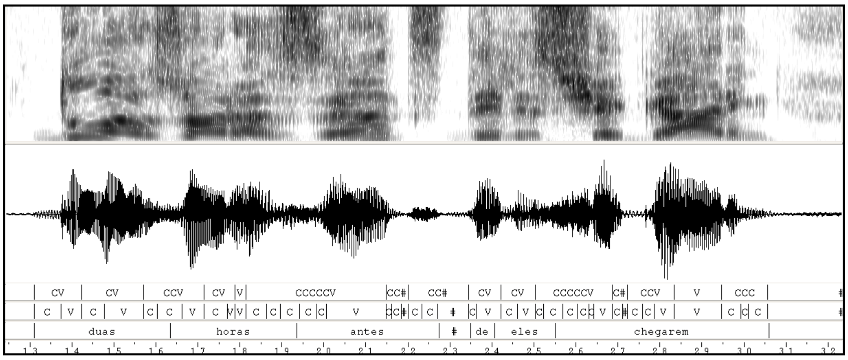

Figure 6. Spectrogram and signal representation for the sentence "na mesma noite **duas horas antes de eles chegarem** uma casa havia sido assaltada na cidade". Transcriptions are: (a) Manual word annotation, (b) automatic segmentation and labeling, (c) pseudo-syllables.

4.4.2 Prosodic coding

Two models are used to separate the long-term and short-term components of prosody. The long-term component characterizes prosodic movements over several pseudo-syllables while the short-term component represents prosodic movements inside a pseudo-syllable. The fundamental frequency processing is divided into two phases, representing the phrase accentuation and the local accentuation, as in Fujisaki's work [13]. The phrase accentuation is used for the long-term model while the local accentuation is used for the short-term model. Fundamental frequency and energy are extracted from the signal using the SNACK Sound toolkit [37].

The long-term coding uses the pseudo-syllable segmentation as a time-base. The coding is described in Figure 7. The "baseline" is a representation of the phrase

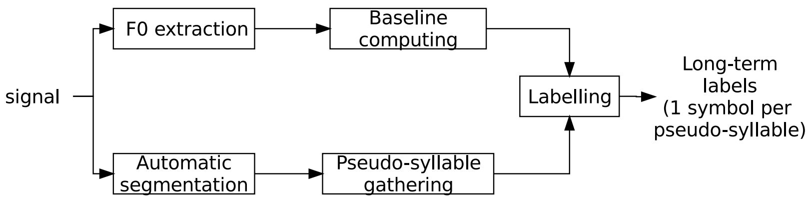

Figure 7. Long-term coding.

accentuation. It is computed by finding all the local minima of the F_0 contour, and linking them. The labels used are U(p), D(own), respectively representing a positive and a negative slope of the baseline, and #(silence or unvoiced).

An example of a resulting baseline curve is displayed on Figure 8. On this example, with one label per pseudo-syllable, the label sequence corresponding to the sentence is:

```
U.U.#.D.D.D.D.#.D.D.#.D.U.U.U
```

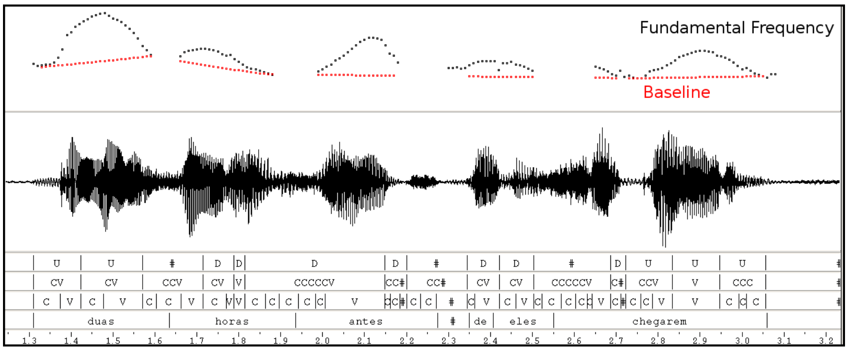

Figure 8. Extraction of the baseline on the same sentence. Previously found local minima are linked with a straight line.

The short-term coding is detailed on Figure 9. The short-term coding use the “C”,

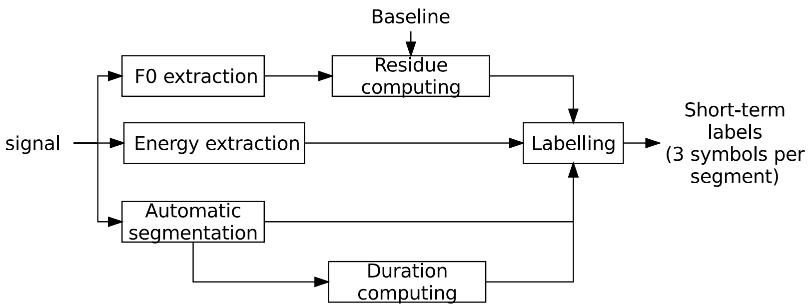

Figure 9. Short-term coding.

“V” and “#” segments as a time base. The local accentuation, named here residue, is represented by the difference between the original F_0 contour and the baseline. This residue is then approximated on each segment by a linear regression. An illustration of this approximation is displayed on Figure 10. The slope of the linear regression is used to label the F_0 movement on each segment, according to three available labels (Up, Down and Silence).

In parallel, the energy curve is computed and also approximated by linear regressions on each segment. The process is the same as the one used for the residue coding. As there is no segment with no energy, only two labels are used: Up and Down.

Duration labels are also computed on the segment units. The labels are assigned considering the mean duration of each kind of segment (vocalic, consonantic or silence). If the segment to be labeled is a vocalic segment with a duration superior

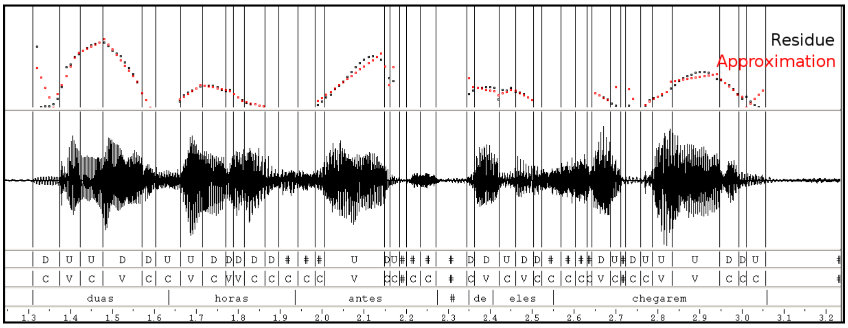

Figure 10. Approximation of the residue on the sentence “na mesma noite **duas horas antes de eles chegarem** uma casa havia sido assaltada na cidade”. The residue is approximated by a linear regression on each segment.

to the mean vocalic segments duration computed on the learning part of the corpus, it is labeled “l” (long), Otherwise, if the current vocalic segment duration is under the mean, the “s” (short) label is used.

These three coding are used conjointly to form the short-term coding. Hence, for each segment, the label is then composed of three symbols. For the example sentence, we have:

```
DUs.UUs.UUs.DDl.DDs.UDs.UUl.DDs.DUs.DDs.DDs.DDs.#Ds.-
#Us.#Us.UDl.DDs.UDs.##s.#Us.#Ds.##l.DUs.DUl.UDs.DDs.-
DDs.#Us.#Us.#Us.#Ds.DUs.UDs.##s.DDs.UUs.UUs.UDl.DDs.-
DDs.UDs
```

4.4.3 Prosodic N-gram Modeling

To model the prosodic variations, we use classical n-gram language modeling provided by the SRI language modeling toolkit [38]. For each system – long- and short-term – each target language is modeled by a n-gram model during the learning procedure. A background model is also learned using data from all languages. During the test phase, the most likely language is picked according to the model (target or background) which provides the maximum likelihood.

This system has been previously applied to different kinds of databases [34], which have shown its ability to model language-specific prosody patterns, especially in the case of studio-recorded read speech. The broadcast news corpora used here contains both prepared and spontaneous speech, and it will be interesting to study how this prosodic system deals with the disfluencies present in spontaneous speech (hesitations, filled pauses, repetitions, fragments, ...).

5 Corpora

Two different corpora have been used for the experiments. The first corpus is used for the language verification experiment, i.e. to test the reliability of the language identification system, especially for Portuguese broadcast news. The second corpus is used for variety identification.

For the language verification experiments, we used the COST 278 corpus, which contains records of several broadcast news from different European countries. Since there were no English recordings in this corpus, and given the fact that English is the most frequent language, next to Portuguese, found in Portuguese broadcast news, we complemented the COST 278 corpus with a part of the HUB4 corpus which contains American broadcast news.

For the variety identification experiments we have used the EP part of the COST 278 corpus and we have recorded several shows from the "Reporter Africa" program on RTP-Africa, and several main evening shows from the Brazilian channel TV Record.

5.1 Language Verification Corpora

5.1.1 The COST 278 Corpus

This corpus was constructed by seven institutions that collaborated in the European COST 278 action on Spoken Language Interaction in Telecommunications [3]. It comprises broadcast news shows in nine languages, namely Dutch (from Belgium, noted BE), European Portuguese (EP), Galician (GA), Czech (CZ), Sloven (SI), Slovakian (SK), Greek (GR), Croatian (HR) and Hungarian (HU) [41].

The first part of Table 1 shows the names of the TV stations, the number of collected shows and the total data size (in minutes).

One important language is missing in this corpus however, as there are no English recordings. That is why we have used English recordings from an other broadcast news corpus.

[3] http://cost278.org/

Table 1
Overview of the COST 278 corpus (top part) complemented with the subset of the HUB4 corpus (bottom).

Code	Country	Language	# of shows	Duration (min)
BE	Belgium	Dutch	6	150
CZ	Czech Rep.	Czech	5	171
GA	Spain	Galician	4	184
GR	Greece	Greek	3	174
HR	Croatia	Croatian	6	166
HU	Hungary	Hungarian	11	166
EP	Portugal	Portuguese	6	190
SI	Slovenia	Sloven	3	151
SK	Slovakia	Slovak	9	165
EN	United States	English	10	328
Total	10	10	66	33h47min

5.1.2 English recordings: (HUB4 1996 Corpus)

The 1996 Broadcast News Speech Corpus contains a total of 104 hours of broadcasts from ABC, CNN and CSPAN television networks and NPR and PRI radio networks with corresponding transcripts. The primary motivation for this collection was to provide training data for the DARPA "HUB4" Project on continuous speech recognition in the broadcast domain.

The speech files are available in a 19 disc training data set with one additional disc of development data and an additional disc of evaluation data. The data includes programs from 11 different programs (recorded from the ABC, CNN, CSPAN and NPR channels). Transcripts have been made of all recordings in this publication, manually time aligned to the phrasal level, annotated to identify boundaries between news stories, speaker turn boundaries and gender information about the speakers.

For the purpose of our language identification studies, we only used the first data CD in order to keep consistency with the amount of data available in the COST 278 corpus. The programs used in our experiments are 10 shows from “ABC Nightline”, with a mean duration of approximately 30 min. The corresponding information is shown in the bottom part of Table 1.

5.2 *Train and test sets*

Train and test sets have been defined for each language. The test set contains one or two shows per language. The remaining shows are used in the train set. Some statistics about the data are displayed in Table 2.

The train set contains a total of 1659 automatically detected speech segments, for a total duration of 16 hours and 12 minutes. The duration per language ranges from 114 minutes to 168 minutes.

Table 2
Description of the train and test sets.

	Train		Test	
Language	# speakers	Total duration (min)	# speakers	Total duration (s)
BE	180	117	58	37
CZ	63	114	94	63
EN	612	221	57	30
GA	380	149	137	75
GR	265	139	122	58
HR	293	138	54	33
HU	261	143	57	28
EP	257	168	39	24
SI	156	102	120	60
SK	225	140	51	27

5.3 *Variety Verification Corpus*

5.3.1 *Reporter Africa*

Reporter Africa is the main news programs from the RTP Africa channel. Each daily show lasts for 30 minutes, with information from reporters in Angola, Cabo Verde, Guinea-Bissau, Mozambique, São Tomé and Príncipe. The anchor speaks European Portuguese. We have recorded 16 shows and labeled the varieties for each of these. The total duration is 480 minutes, but we have excluded the EP speakers, foreign speakers and also the few speakers for which the human annotators could not distinguish the country of birth, being only able to tell they were from Africa.

The number of speakers and the duration (in minutes) for each African variety is shown in Table 3.

Table 3
Portuguese varieties recordings from "Reporter África".

Code	Country	# speakers	duration (minutes)
AN	Angola	86	41
CV	Cape Verde	81	37
GB	Guinea-Bissau	86	43
MO	Mozambique	69	44
ST	São Tomé and Principe	70	29
Total	5	430	194

5.3.2 Brazilian data

The Brazilian recordings come from the main news show of the TV Record Brazilian channel. We have recorded 6 40-minutes shows (including publicity). After pre-processing, we have a total of 190 minutes of Brazilian speech data, with 404 automatically detected speakers.

5.3.3 Train and Test sets

Given the relative low volume of data for each variety, we have decided to carry on experiments using a cross-validation procedure. First, one speaker is selected for testing. All the remaining data is used for learning the variety models. After the test is achieved, a new speaker is used for testing. This procedure is iterated until all the speakers of the corpus have been used for testing. We will then have a total of 1144 tests, with the 430 speakers of African varieties, the 404 Brazilian speakers, and the 310 European Portuguese speakers from the COST 278 corpus. The scenario is not ideal as speaker clustering is done on a file-by-file basis and some speakers may be the same in the training and test sets, even excluding the very frequent speakers (e.g. anchormen). Getting more data from each variety (namely from Africa) is thus one of our current goals.

6 Evaluation of the Pre-Processing Module

Because of their influence on the LID results, it is important to report the results of some of the blocks of the pre-processing module. The evaluation was conducted using the COST 278 multilingual corpus.

The results for the acoustic change detector are displayed in Table 4 in terms of Recall (% of detected acoustic change points), Precision (% of detected points which are genuine change points), and F-measure (defined as $(2 * Recall * Precision)/(Recall + Precision)$.

Table 4
Evaluation of the acoustic change detector.

Recall (%)	Precision (%)	F-measure (%)
78.9	65.5	70.9

The performances of the Speech/Non speech and Gender classifiers are shown in Table 5.

Table 5
Performances of the Speech/Non speech and Gender classifiers (in terms of % of correctly identified frames).

	Speech	Non Speech	Accuracy
Speech/Non Speech	97.5	70.6	95.6
	Male	Female	Accuracy
Gender	96.7	90.2	94.5

Figure 11 shows the results obtained in term of accuracy for the Speech/Non speech detector for each of the languages of the COST 278 corpus.

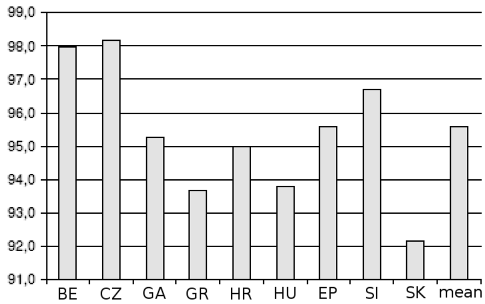

Figure 11. Accuracy results for the Speech/Non speech detector on the languages of the COST 278 corpus.

Table 6 shows the performances of the speaker clustering system. The results are given in terms of Q-measure and Diarization Error Rate (DER). The Q-measure is defined as the geometrical mean of the percentage of cluster frames belonging to the correct speaker and the percentage of speaker frames labeled with the correct cluster, and the DER is the percentage of frames with a incorrect cluster-speaker correspondence. As one speaker is often divided in several clusters, the performance in terms of DER is not very high, but this type of error does not affect the LID system.

Table 6
Evaluation of the speaker clustering.

Q (%)	Q_{map} (%)	DER (%)
68.1	87.8	31.6

Experiments on the Portuguese part of the COST 278 corpus have shown that using this pre-processing module has very little influence on the performance of the speech recognizer as compared to using manual speaker segmentation [28].

7 Language Verification Experiments

The verification framework was used in the recent NIST language recognition evaluation campaigns (see http://www.nist.gov/speech/tests/lang/). The task is the detection of a given language: given a test segment of speech and a target language, the system must determine whether or not the speech is from the target language.

The performance of the system is characterized by its miss and false alarms probabilities. Results displayed in this chapter are computed with the NIST DET curve software [4] . The overall performance is qualified by the Equal Error Rate (EER) for each system, which is given when the false alarm rate is balanced with the missed detection rate.

7.1 Results

In the language verification database, we have a total number of 789 test speakers. Considering that we test the detection for all the 10 languages, we have a total of 7890 language verification tests, with 789 target trials and 7101 non-target trials. The results are displayed in the Figure 12. The fusion (weighted addition of log-likelihoods) is achieved using empirically determined weights.

The best results are unsurprisingly given by the PRLM system. The GMM-MFCC system gives the second best results, followed by the prosodic short-term system.

[4] http://www.nist.gov/speech/tools/

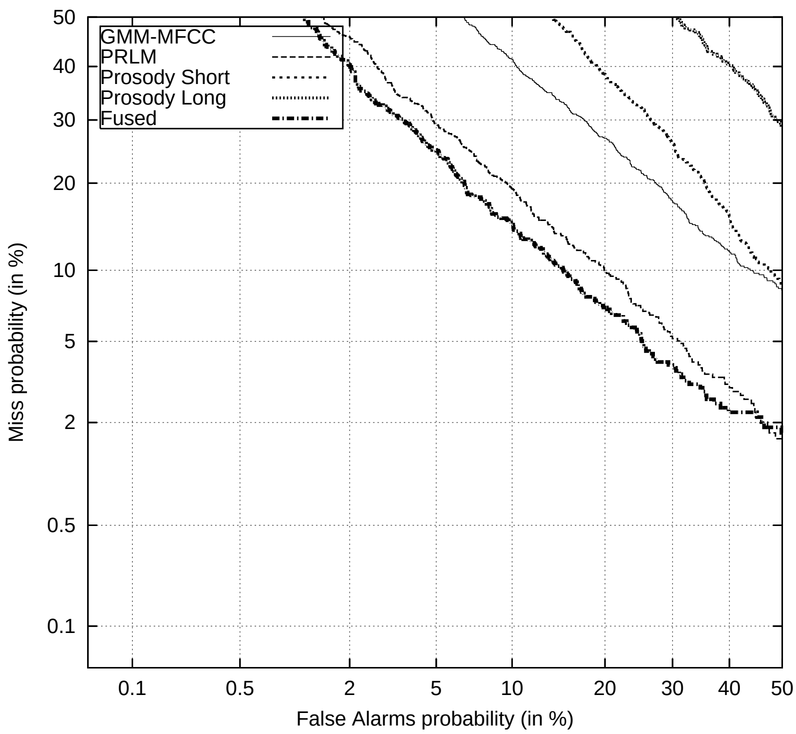

Figure 12. DET curve obtained using all the systems.

The DET curve obtained using the fused system is also displayed in Figure 12.

Results are given in terms of Equal Error Rate (EER) in Table 7, for each system and each language. In this Table, different thresholds are used for computing the EER for each language. The EER is meant to give a point of the performance of the system when balanced between false alarms and false rejections – that is to say on the diagonal of the graph displayed in Figure 12.

Table 7
Results per language in terms of % EER.

Language	BE	CZ	EN	GA	GR	HR	HU	EP	SI	SK	ALL
Fused (% EER)	3.6	13.8	5.3	6.6	19.7	9.1	5.2	2.5	19.0	18.1	12.4

The overall performance of the fused system is 12.4% EER. There are great differences in performance considering the language, the worst performance is obtained for Greek with 19.7 % EER, while the best performance is 2.5 % EER for Portuguese. This is not at all surprising, given the use of the Portuguese phone recognizer in the PRLM system. The fused system also achieves good performance for Belgian Dutch (3.6% EER), English (5.3% EER) and Hungarian (5.2% EER).

The Tables 8 and 9 show respectively the number of false alarms and missed detections over all languages and, given the focus of our work, for Portuguese. These results are computed using the fused system.

Table 8
False detections for European Portuguese.

Language	# false alarms	# non-targets trials	% EER
EP	**19**	**750**	**2.5**
ALL	888	7101	12.5

Table 9
Missed detections for European Portuguese.

Language	# Missed detections	# targets trials	% EER
EP	**1**	**39**	**2.6**
ALL	98	789	12.4

The Portuguese false alarms are distributed across the different languages: 1 from English, 9 from Galician, 5 from Greek, 3 from Sloven, and 1 from Slovak. As Galician and Portuguese are closely related, it is not surprising to find that some Galician speakers are identified as Portuguese. All the errors are linked either to bad acoustic conditions (e.g. live sports reports) or very short test segments (e.g. duration under 5 seconds).

The only missed detection error for the Portuguese verification system appears on a 2 second segment, which is in fact a music segment in English, wrongly labeled as Portuguese.

As a result from this analysis, we can hypothesize that the acoustic environment and the short length of the test segments, combined with pre-processing errors, are the main factors that lead the system to generate errors, at least for the Portuguese language. This behavior seems however to be the same for all languages. Since the detection of the acoustic environment is one of the tasks of the pre-processing module (see section 4.1), we will take advantage of the current work towards its improvement. Concerning the duration of the test segments, this issue is addressed in the next section.

7.2 Impact of the test segment duration

As the duration of the test segment varies greatly, we have investigated how the performance of the system increases when discarding very short segments. These experiments show how the different systems work with segments with minimum length of 10s, 20s, and 30s. The number of test segments reduces when only long segments are selected:

- no length constraint : 789 test segments,

- segments >10s : 618 test segments,
- segments >20s : 394 test segments,
- segments >30s : 272 test segments.

Table 10
Results for all the systems in terms of % EER depending on the minimum duration of the test segments.

Min. duration	0s	10s	20s	30s
Fused system (% EER)	12.4	9.3	6.6	5.8

Table 10 shows that, as expected, the performance of the system improves when using longer segments. The improvement is clearly significant when selecting segments of duration over 30s: the EER becomes 5.8% instead of 12.4% when no length constraint is applied. This is illustrated by the DET-curve displayed on Figure 13.

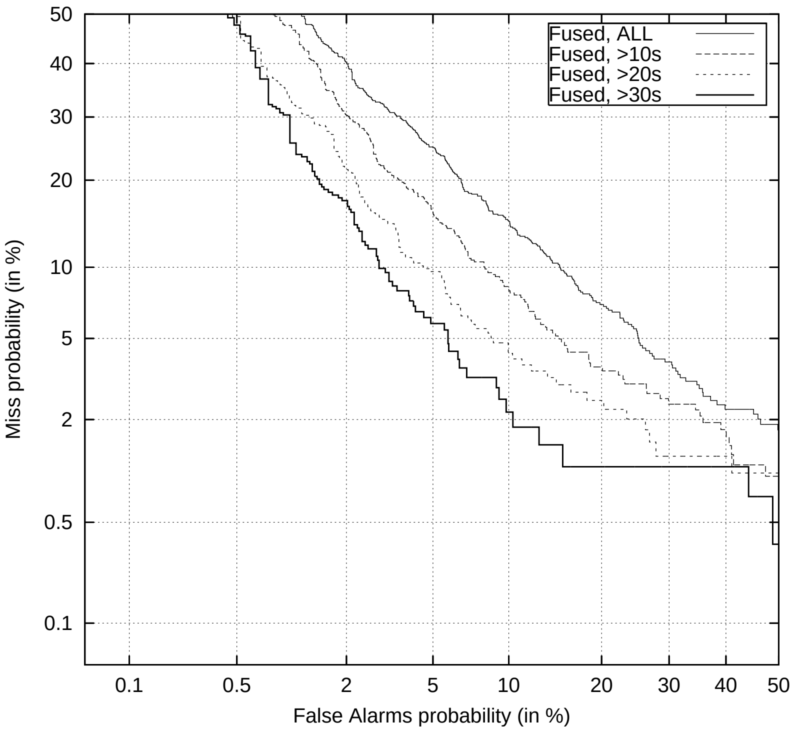

Figure 13. DET curve obtained using the fused system.

Selecting segments that have enough information for identifying the language is clearly needed to achieve better performance.

The same type of analysis is shown in particular for the Portuguese language verifier in Table 11. One can observe that the system does not make any errors for files over 20 seconds, and that the error rate is only 0.2% for files over 10 seconds. As the error rate seems sufficiently low for the Portuguese verification task, the next step is to investigate how this system behaves when trying to identify the different varieties of Portuguese – European, Brazilian, and African.

Table 11
Performances in terms of % EER of the European Portuguese language verifier depending on the test segments duration.

Min. duration	0s	10s	20s	30s
# test segments	789	618	394	272
EP Verif. (% EER)	2.5	0.2	0.00	0.00

8 Behavior of the language verifier with other varieties of Portuguese

The aim of this experiment is to investigate how the Portuguese language verifier behaves when confronted with data from all the varieties of Portuguese. In the experiments described above, we only have considered European Portuguese. For this experiment, we used all the data described in section 5.3. What is expected is that this data should be recognized as Portuguese (as opposed to the other languages: English, Dutch, etc.), leaving the possibility of a second classification phase designed to detect the Portuguese variety.

Table 12 shows the results obtained for all the segments durations. When not selecting a minimum length, there is a total of 1144 segments to test. The number of test segments becomes 864 when considering segments of duration superior to 10 seconds, 594 for segments superior to 20 seconds, and 402 for segments superior to 30 seconds.

Table 12
Results for the system in terms of % EER depending on the minimum duration of the test segments.

Min. duration	0s	10s	20s	30s
# test segments	1144	864	594	402
Fused system (% EER)	16.1	13.5	11.5	7.2

In Table 13, results are detailed for each variety, considering only the segments of duration superior to 30 seconds.

The variety which is less recognized as Portuguese is Brazilian Portuguese, which

Table 13
Results per variety using the fused system (segments > 30s).

Variety	AN	BP	CV	EP	GB	MO	ST
# Missed	0	31	0	0	0	0	0

is responsible for all the errors of the system. A closer look at the errors allows us to see that most occur during the headlines or weather forecasts, which contain loud background music.

9 Discrimination between Portuguese varieties

9.1 Global discrimination

This experiment aims at testing the system while trying to discriminate between all the Portuguese varieties at the same time. The tests are made according to the cross-validation procedure described in section 5.3. The European Portuguese corpus used for this experiment is from the COST 278 corpus described in section 5. The global variety identification results are shown in Table 14.

Table 14
Identification of Portuguese varieties - confusion matrix produced by the fused system. (%)

	AN	BP	CV	EP	GB	MO	ST
AN	31.2	26.0	1.3	37.7	0.0	2.6	1.3
BP	0.0	97.9	0.5	1.6	0.0	0.0	0.0
CV	0.0	5.3	36.0	53.3	0.0	5.3	0.0
EP	0.7	1.7	0.7	88.4	7.6	1.0	0.0
GB	6.3	11.4	3.8	74.7	0.0	3.8	0.0
MO	7.5	16.4	0.0	40.3	0.0	35.8	0.0
ST	9.4	17.2	4.7	53.1	0.0	7.8	7.8

The global identification rate is 69.0%. The best recognised variety is Brazilian (97.9%), followed by European Portuguese (88.4%). The varieties from Guinea-Bissau and São Tomé and Principe are the worst recognised (0% and 7.8%).

The lack of data for the African varieties – as compared with European and Brazilian Portuguese – may explain the poor performance achieved on this data by the system. In order to provide a more balanced experiment, we adress in the next section the identification of broad varieties by regrouping all the African varieties into one class. Using the grouping, we have almost the same amount of data for each broad variety (194 minutes for AP, 190 minutes for BP and 194 minutes for the EP part of the COST 278 corpus). The discrimination between African varieties will then be described in the following section.

9.2 Discrimination between African, Brazilian and European Portuguese

For the purpose of this experiment, we have gathered data from all the African Portuguese-speaking countries into a broad variety: African Portuguese. Thus, the aim of the designed system is to verify is the test speaker speaks Brazilian, European or African Portuguese.

Table 15
Identification of Portuguese varieties - Confusion matrix using only 3 broad classes (African, Brazilian and European Portuguese).

	AP	BP	EP
AP	93.0	6.3	0.7
BP	5.5	94.5	0.0
EP	2.3	0.6	97.1

The designed system performs quite well on this data, with a global identification rate of 94.7%. Detailed results (Table 15) show that the best identified variety is European Portuguese (97.1%). This result is obtained using the fused system. It is however noticeable that the prosodic system alone achieves an identification rate of more than 77%.

9.3 Discrimination between African varieties

After identifying that a speaker speaks African Portuguese, the following experiment aims at finding which African Portuguese variety is actually spoken.

The global identification rate (see Table 16) is 60.1%. The most clearly identified varieties are Portuguese from Guinea-Bissau (73.7%) and Angola (71.2%). The weakest identification rate is for São Tomé and Príncipe.

Table 16
Identification of African Portuguese varieties - Confusion matrix produced by the fused system (correct=60.1%).

	AN	CV	GB	MO	ST
AN	71.2	6.8	10.9	6.8	4.1
CV	2.8	60.6	22.5	9.8	4.2
GB	9.2	5.2	73.7	9.2	2.6
MO	20.3	3.1	14.0	59.4	3.1
ST	32.3	14.5	11.3	11.3	30.6

9.4 Human benchmark experiment

In order to compare the performance of our automatic variety identification system with a manual one, we conducted a human benchmark. For this purpose, we have selected 8 stimuli from each of the 7 varieties. In this selection, we avoided sentences that could give an indication either by lexical, syntactical or semantical cues of the origin of the speaker. That is, we avoided the mention of locations, politicians, political parties, etc.. We also avoided sentences with clitics, since the Brazilian origin would be very noticeable, and sentences where the lack of number agreement would make the African origin too noticeable. In this way, the human benchmark test was made in conditions as close as possible to the ones of our automatic variety identification system. The sentences (or segments from sentences) ranged in duration between 1.6 and 23.4 seconds. Most of the sentences were extracted from spontaneous speech (64%), in order to avoid easily identifiable journalists or politicians. In addition to the 8 sentences, the participants were asked to identify the variety of 2 words (also extracted from sentences). The total duration of all stimuli was 8.5 minutes. Participants were asked to classify each stimulus as one of the 7 varieties, but they also had an option to mark it as African Portuguese (AP). In very few cases they forgot to (or could not) mark their preference (no answer - NA).

The test involved 65 participants. 44 participants were Portuguese, 7 were from Brazil and 14 from Africa (8 from Angola, 4 from Cape Verde and 2 from Mozambique).

Table 17 shows the confusion matrix results of this test, with a dividing line between sentences (top part) and words (bottom part). The results very clearly show that, as in the automatic test, Brazilian Portuguese is the least confusable variety. They also show that European Portuguese is next and that African varieties are easily confused with each other. Among these varieties, ST was the hardest to identify. The results with words were naturally inferior, except for BP, showing a greater tendency towards classifying African varieties as AP.

Table 17
Human benchmark results (% of correct identification). The top part shows the results with sentences and the bottom part shows the results with words.

Variety	AN	BP	CV	EP	GB	MO	ST	AP	NA
AN	20.0	0.6	7.5	0.0	7.3	9.2	8.1	47.3	0.0
BP	0.0	99.2	0.4	0.0	0.0	0.0	0.2	0.2	0.0
CV	11.0	0.4	16.5	4.8	4.0	10.4	6.7	45.8	0.4
EP	1.9	0.6	1.3	88.7	0.4	1.0	0.8	5.4	0.0
GB	17.7	0.2	8.3	2.1	10.0	8.7	7.7	45.2	0.2
MO	13.7	0.2	5.4	1.5	7.7	14.6	9.4	47.1	0.4
ST	14.4	1.2	10.4	2.5	8.1	10.2	9.2	43.8	0.2
AN	20.8	0.0	2.3	0.8	3.1	6.9	5.4	60.0	0.8
BP	0.0	99.2	0.0	0.0	0.0	0.0	0.0	0.0	0.8
CV	4.6	1.5	12.3	3.1	6.2	3.1	4.6	63.8	0.8
EP	1.5	1.5	4.6	73.8	0.0	3.8	1.5	13.1	0.0
GB	10.0	0.0	3.8	3.8	6.2	11.5	6.9	57.7	0.0
MO	10.0	0.0	4.6	7.7	8.5	7.7	5.4	56.2	0.0
ST	12.3	0.0	5.4	11.5	2.3	7.7	4.6	55.4	0.8

It was interesting to notice that practically all Portuguese participants correctly identified BP and (although not so clearly) EP sentences, and most could correctly identify African varieties as such but, even if they have some suspicion about the African country of origin, namely if they have lived there, they were often reluctant to discriminate. Some Brazilian participants had no familiarity at all with African varieties, tending to confuse them with EP. Most African participants correctly identified BP and EP varieties, but they also tried to discriminate between African varieties more often. Their general opinion was that identifying African varieties in BN was much more difficult than identifying the varieties of the African people they meet everyday, most probably because in BN, many speakers (reporters, politicians and people involved in cultural events) have a higher level of education and/or familiarity with EP.

If these results are analyzed using only three broad classes (AP, BP and EP), as shown in Table 18, the average ratio of correct identification is 96.2% for sentences and 91.8% for words.

Just for comparison purposes, we have also run an experiment aimed at investigat-

Table 18
Human benchmark results with only 3 broad classes. The top part shows the results with sentences (correct=96.2%) and the bottom part shows the results with words (correct=91.8%).

Variety	AP	BP	EP	NA
AP	97.1	0.5	2.2	0.2
BP	0.8	99.2	0.0	0.0
EP	10.8	0.6	88.7	0.0
AP	93.8	0.3	5.4	0.5
BP	0.0	99.2	0.0	0.8
EP	24.6	1.5	73.8	0.0

ing the behavior of the automatic variety identification system with these stimuli. The number of files is too small to get any significant results, and some of the files were too short, but still the automatic 3-class system yielded reasonably good results (above 70%).

9.5 Automatic speech recognition experiments

It is also interesting to relate the results of the automatic/human variety identification tests with the results obtained with an automatic speech recognition system trained for Broadcast News in EP. The acoustic models of this system have already been described. The vocabulary includes around 57k words. The lexicon includes multiple pronunciations, totaling 65k entries. The corresponding out-of-vocabulary (OOV) rate is 1.4%. The language model, which is a 4-gram backoff model, was created by interpolating a 4-gram newspaper text language model built from over 604M words with a 3-gram model built on around 532k words of manually transcribed broadcast news ($\approx$ 50 hours). The language models were smoothed using Knesser-Ney discounting and entropy pruning. The perplexity obtained in a development set is 112.9.

Table 19 shows the ASR results in terms of word error rate (WER), obtained using all the training/test material of our accent id system. The best performance was obtained for EP, obviously. The fact that the acoustic phones used in the PRLM module were the same as in the ASR module justifies the best performance of PRLM for this variety. The percentage of spontaneous speech in this subset is relatively low, which may also account for the low WER obtained. [5] The worst performance was obtained for BP, a fact that was also expected given that it was so easily distinguish-

[5] In other sets with a percentage of spontaneous speech closer to 40%, the WER goes up to 23.5%.

able from EP, both manually and automatically. Intermediate results were obtained for all African varieties, with WER values ranging between 30.1% (AN) and 37.8% (ST). These values are closer to the ones obtained for EP than to the ones obtained for BP and hence are in agreement with the fact that African varieties are more easily confused with EP than with BP. The OOV rate for Brazilian and African varieties is not significantly higher than the one obtained for EP (1.8% for BP and 2.0% for AP) thus not being responsible for the large performance degradation.

Table 19
Word error rate results obtained on the multi-variety corpus by an EP-trained ASR system.

Variety	AN	BP	CV	EP	GB	MO	ST
% WER	30.1	62.3	31.1	15.2	35.4	34.5	37.8

10 Conclusions and Future Work

The first part of this paper described a language verification system for Broadcast News. The system is composed of three modules used to model language discriminative features: phonotactics, acoustics and prosody. Over all the 10 languages of the multilingual BN corpus we have used, the average performance of the fused system is 12.4% EER. The comparison with other systems is not straightforward, since there is not so much reported work on Broadcast News data, and none with as many languages as we have used. The EER obtained with the fused system on the "segments over 30s" condition (5.8%) may be compared to the best results obtained on the NIST 2005 data (4.2% EER). The corpora used in both evaluations are, however, quite different. The NIST 2005 data is telephone speech, which is likely to have worse quality than broadcast news, but does not include so much diversity in terms of acoustic conditions, prepared and spontaneous speech, etc.. In fact, one of the approaches we are currently investigating in order to improve our system is to take into account these different acoustic conditions. Another difference between the two corpora lies in the constraints on the homogeneity of the segments: in the NIST corpus there is exactly one speaker per file, whereas in our broadcast news corpus, automatic speaker clustering is adopted, thus potentially generating some errors.

Not surprisingly, since the phonotactics module used the acoustic models of an ASR system trained for European Portuguese, the best performance of our language verification system was achieved for this language (2.5% EER). A further analysis of the performance of the system revealed that the false alarms errors occurred mainly while misidentifying Galician speakers, and the missed detection errors appeared only on short files, some of them with much background noise or non-speech segments, erroneously classified as speech by an automatic audio pre-processing system. When tested over segments of duration above 10s, the equal error rate drops to 0.2% ERR, and no errors were observed when considering segments above 20s. Hence we may consider that the language verification is robust enough to be integrated in our Broadcast News recognition system in order to exclude non-Portuguese speech segments, which was the real goal of this work.

A further experiment was conducted involving a different corpus which includes BN data from other varieties of Portuguese, namely the ones spoken in Brazil and in African countries with Portuguese as official language. In this experiment, the error rate is above the language verification error rate mentioned above (7.2% for the 30 second test segments), but most errors seem again to come from the bad acoustic conditions of the test excerpts, which often contain loud background music (typically the jingles that mark headlines or weather forecast news).

This experiment showed that the verification system can cope with other varieties of Portuguese. However, some of these varieties can cause a severe degradation of the performance of the recognizer. Hence, the second part of this work was devoted to the study of an accent identification system for Portuguese, using this multi-variety BN corpus.

Our accent identification system achieved an average correct identification ratio of 69.0%. The least confusable variety was by far BP (97.9% correct identification). EP was next. African varieties were the hardest to discriminate. That led us into trying to build a system with only 3 broad classes: AP, BP or EP. The average ratio achieved by this system was 94.7%.

The results of these experiments were compared with the ones of a human benchmark test, which basically revealed a very good capacity for detecting BP and, although not so easily, EP, and similar difficulties in discriminating African varieties, although they could also be easily identified as such. The average 3-class identification ratio was 96.2% for sentences.

Finally, the results were also discussed in view of the performance of an EP-trained speech recognition system when confronted with other varieties. Given the strong degradation namely for BP (WER=62.3%), the adaptation of the models of our EP-trained recognizer to these varieties is one of the topics we are currently pursuing.

There are many ways in which the above described language/variety identification methods can be improved. For instance:

- The PRLM can be improved by adding other languages phones to the phone recognizer, or by using several language-specific phone recognizers. Another point can be considering phone lattices, as proposed initially in [15] and used in the MIT system on the NIST 2005 language recognition data.
- The acoustic system can be improved by using different kinds of models: recent research has shown an interest in SVMs (especially for the language verification

framework for which they are more suited). ANNs can also be investigated.
- The prosodic system can be modified using a better definition of a pseudo-syllable, by taking into account the different types of vowels and consonants. An important issue is to take into account the variations in terms of speaking rate that can occur in different speaking styles.
- The fusion procedure can be much more sophisticated. For instance, one can implement a backend classifier using either Neural Networks or Fuzzy Logic algorithms.

11 Acknowledgments

The authors would like to thank our colleagues Hugo Meinedo and Ernesto de Andrade for helpful comments. This work was partially funded by FCT under the post-doc scholarship SFRH/BPD/22032/2005 and project WFST (POSI/PLP/47175/2002) and also by the European project Vidi-Video. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

References

[1] M. B. Abaurre, E. G. Pagotto, Nasalização no Português do Brasil, in: I. Koch (ed.), Gramática do Português Falado, vol. VI, Editora da Unicamp/FAPESP, Campinas SP, 1996, pp. 495–526.

[2] R. André-Obrecht, A new statistical approach for automatic speech segmentation, IEEE Transactions on Acoustics, Speech and Signal Processing 36 (1) (1988) 29–40.

[3] P. Barbosa, E. Albano, Brazilian Portuguese - illustrations of the IPA, Journal of the International Phonetic Association 34 (2) (2004) 227–232.

[4] K. Berkling, M. Zissman, J. Vonwiller, C. Cleirigh, Improving accent identification through knowledge of english syllable structure, in: ICSLP'98, Sydney, Australia, 1998.

[5] D. Callou, Y. Leite, Iniciação à Fonética e Fonologia, Jorge Zahar Editor, Rio de Janeiro, 1990.

[6] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, P. Torres-Carrasquillo, Advanced language recognition using cepstra and phonotactics: MIT-LL system performance on the NIST 2005 LRE, in: Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop, 2006.

[7] E. Campione, J. Véronis, A multilingual prosodic database, in: International Conference on Spoken Language Processing, Sidney, 1998, http://www.lpl.univ-aix.fr/projects/multext.

[8] T. Chen, C. Huang, E. Chang, J. Wang, Automatic accent identification using gaussian mixture models, in: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2001.

[9] R. M. Dauer, Stress-timing and syllable-timing reanalysed, Journal of Phonetics 11 (1983) 51–62.

[10] F. Fernandes, Subject localization constraints in BP and EP, Ph.D. thesis, Univ. Estadual Campinas (2007).

[11] S. Frota, M. Vigario, F. Martins, Language discrimination and rhythm classes: evidence from portuguese, in: Speech Prosody, Aix-en-Provence, France, 2002.

[12] S. Frota, M. Vigário, On the correlates of rhythmic distinctions: The European / Brazilian Portuguese case, Probus 13.

[13] H. Fujisaki, Prosody, information and modeling - with emphasis on tonal features of speech, in: ISCA Workshop on Spoken Language Processing, Mumbai, India, 2003.

[14] P. Fung, L. Kat, Fast accent identification and accented speech recognition, in: ICASSP'1999, Phoenix, Arizona, 1999.

[15] J.-L. Gauvain, A. Messaoudi, H. Schwenk, Language recognition using phone lattices, in: INTERSPEECH'2004, Jeju, Korea, 2004.

[16] R. Huang, J. H. Hansen, Gaussian mixture selection and data selection for unsupervised spanish dialect classification, in: INTERSPEECH'2006, Pittsburgh, Pennsylvania, 2006.

[17] A. Ikeno, J. H. Hansen, The role of prosody in the perception of us native english accents, in: INTERSPEECH'2006, Pittsburgh, Pennsylvania, 2006.

[18] S. Kitazawa, Periodicity of Japanese accent in continuous speech, in: Speech Prosody, Aix en Provence, France, 2002.

[19] M. Komatsu, T. Arai, T. Sugawara, Perceptual discrimination of prosodic types, in: Speech Prosody, Nara, Japan, 2004.

[20] A. Lacerda, N. Rossi, Particularidades fonéticas do comportamento elocucional da fala do Rio de Janeiro (em confronto com o Português normal de Portugal), Revista do Laboratório de Fonética Experimental da Faculdade de Letras da Universidade de Coimbra IV (1958) 5–102.

[21] Y. Leite, D. Callou, J. Morais, Neutralização e realização fonética: a harmonia vocálica no Português do Brasil, in: Actas do Congresso Internacional sobre o Português, vol. III, Lisboa, 1996.

[22] J. Li, S. Yaman, C.-H. Lee, B. Ma, R. Tong, D. Zhu, H. Li, Language recognition based on score distribution feature vectors and discriminative classifier fusion, in: Proceedings of Odyssey06: The Speaker and Language Recognition Workshop, 2006.

[23] M. Lincoln, S. Cox, S. Ringland, A comparison of two unsupervised approaches to accent identification, in: ICSLP'98, Sydney, Australia, 1998.

[24] P. Matejka, L. Burget, P. Schwarz, J. Cernocký, Brno university of technology system for nist 2005 language recognition evaluation, in: Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop, 2006.

[25] M. H. Mateus, E. d'Andrade, The Phonology of Portuguese, Oxford University Press, Oxford, 2000.

[26] H. Meinedo, D. Caseiro, J. Neto, I. Trancoso, Audimus.media: a broadcast news speech recognition system for the european portuguese language, in: PROPOR'2003 - 6th International Workshop on Computational Processing of the Portuguese Language, 2003.

[27] H. Meinedo, J. Neto, Audio segmentation, classification and clustering in a broadcast news task, in: ICASSP'2003, Hong Kong, 2003.

[28] H. Meinedo, J. Neto, A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models, in: INTERSPEECH'2005, 2005.

[29] S. Parkinson, Portuguese, in: M. H. . N. V. (eds.) (ed.), The Romance Languages, 1988, pp. 131–169.

[30] F. Pellegrino, Une approche phonétique en identification automatique des langues: la modélisation acoustique des systèmes vocaliques, Ph.D. thesis, Université Paul Sabatier, Toulouse, France (1998).

[31] F. Pellegrino, R. André-Obrecht, Vocalic system modeling: A VQ approach, in: IEEE Digital Signal Processing, Santorini, 1997.

[32] J.-L. Rouas, Caractérisation et identification automatique des langues, Ph.D. thesis, Univ. Toulouse 3, France (2005).

[33] J.-L. Rouas, Modeling long and short-term prosody for language identification, in: INTERSPEECH'2005, Lisboa, Portugal, 2005.

[34] J.-L. Rouas, Automatic prosodic variations modelling for language and dialect discrimination, IEEE Transactions on Audio, Speech and Language Processing.

[35] J.-L. Rouas, M. Barkat-Defradas, F. Pellegrino, R. Hamdi, Identification automatique des parlers arabes par la prosodie, in: Journées d'Etude de la Parole, 2006.

[36] T. Schultz, Q. Jin, K. Laskowski, A. Tribble, A. Waibel, Speaker, accent, and language identification using multilingual phone strings, in: HLT'2002, San Diego, California, 2002.

[37] K. Sjölander, The snack sound toolkit.
URL http://www.speech.kth.se/snack/

[38] A. Stolcke, SRILM - an extensible language modeling toolkit, in: INTERSPEECH'2002, Denver, Colorado, 2002.
URL http://www.speech.sri.com/projects/srilm/

[39] P. A. Torres-Carrasquillo, T. P. Gleason, D. A. Reynolds, Dialect identification using gaussian mixture models, in: Odyssey: The Speaker and Language Recognition Workshop, Toledo, Spain, 2004.

[40] W.-H. Tsai, W.-W. Chang, Discriminative training of gaussian mixture bigram models with application to Chinese dialect identification, Speech Communication 36 (3-4) (2002) 317–326.

[41] A. Vandecatseye, J.-P. Martens, J. Neto, H. Meinedo, C. Garcia-Mateo, J. Dieguez, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A. Cizmar, H. Papageorgiou, C. Alexandris, The COST 278 pan-european broadcast news database, in: LREC'2004, Lisbon, 2004.

[42] B. Vieru-Dimulescu, P. B. de Mareüil, Perceptual identification and phonetic analysis of 6 foreign accents in french, in: INTERSPEECH'2006, Pittsburgh, USA, 2006.

[43] T. Wu, D. V. Compernolle, J. Duchateau, Q. Yang, J.-P. Martens, Improving the discrimination between native accents when recorded over different channels, in: INTERSPEECH'2005, Lisbon, Portugal, 2006.

[44] B. Yin, E. Ambikairajah, F. Chen, Combining cepstral and prosodic features in language identification, in: International Conference on Pattern Recognition, 2006.

[45] Y. Zheng, R. Sproat, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, S. Yoon, Accent detection and speech recognition for shanghai-accented mandarin, in: INTERSPEECH'2005, Lisbon, Portugal, 2006.

[46] M. A. Zissman, K. M. Berkling, Automatic language identification, Speech Communication 35 (1-2) (2001) 115–124.

Cover Letter

Dear Sirs,

Please find enclosed our submission to the Special Issue on Speech Technologies for Iberian Languages.

Title: Language and Variety Verification on Broadcast News for Portuguese

Authors:

Author 1: Jean-Luc Rouas, INESC-ID / INRETS
Address: INRETS Electronic, Waves and Signal Processing Research Laboratory
for Transport, 20 Rue Élisée Reclus, 59650 Villeneuve d'Ascq, France
Email: rouas@inrets.fr

Author 2: Isabel Trancoso, INESC-ID / IST (Corresponding author)
Address: INESC-ID, Spoken Language Systems Laboratory (L2F)
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
Email: Isabel.Trancoso@inesc-id.pt

Author 3: Ceu Viana, CLUL
Centro de Linguistica da Universidade de Lisboa,
Av. Prof. Gama Pinto, 2, 1649-003 Lisboa - Portugal
Email: mcv@clul.ul.pt

Author 4: Monica Abreu, INESC-ID/IST
Address: INESC-ID, Spoken Language Systems Laboratory (L2F)
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
Email: monica7abreu@gmail.com

Yours sincerely,
Isabel Trancoso