Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2008) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Language and variety verification on broadcast news for Portuguese

Jean-Luc Rouas [a,b], Isabel Trancoso [a,*], Céu Viana [c], Mónica Abreu [a]

[a] INESC-ID, Spoken Language Systems Laboratory (L2F), R. Alves Redol, 9, 1000-029 Lisboa, Portugal
[b] INRETS Electronic, Waves and Signal Processing Research Laboratory for Transport, 20 Rue Élisée Reclus, 59650 Villeneuve d'Ascq, France
[c] Centro de Linguística da Universidade de Lisboa, Av. Prof. Gama Pinto, 2, 1649-003 Lisboa, Portugal

Received 14 June 2007; received in revised form 8 May 2008; accepted 13 May 2008

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 five 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 83.9%, when considering only the three broad varieties, and the best results were obtained for Brazilian Portuguese, also the variety that proved easiest to identify in perceptual experiments. The identification rate between African varieties themselves is relatively low, a fact that was also observed in the perceptual experiments.

Keywords: Language verification; Portuguese varieties

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

* Corresponding author. Tel.: +351 1 310 0268; fax: +351 1 314 5843.
E-mail addresses: rouas@inrets.fr (J.-L. Rouas), Isabel.Trancoso@inesc-id.pt (I. Trancoso), mcv@clul.ul.pt (C. Viana), monica7abreu@gmail.com (M. Abreu).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.05.006

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 6, focusing mainly on the performances of the Portuguese language verifier. In Section 7, 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 8.

1. Introduction

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

In (Zissman and Berkling, 2001), four kinds of cues are described:

- Phonology: the phoneme sets used in different languages differ, even though many languages 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 (Yin et al., 2006).

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 experiments.

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 are a very good ground for estimating language verification system performances, some of the best performing systems of the 2005 evaluation are briefly described below (see www.nist.gov/speech/tests/lang/ for a complete list of participating organizations).

The seven languages used in the NIST 2005 Evaluation were the following: English (American and Indian), Hindi, Japanese, Korean, Mandarin (Mainland and Taiwan), Spanish (Mexican) and Tamil. 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 Matejka et al., 2005) uses a GMM-based acoustic system, with discriminatively trained models, combined with a Neural Network-based PPRLM system. Combining the scores of these approaches with a weighted addition of the log-likelihoods gives an overall equal error rate of 5.0% on 30-second excerpts.

The system submitted by the Georgia Institute of Technology and the Infocomm Institute uses a fusion of two approaches (Li et al., 2006). The first one is a classical PPRLM. The second approach 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 six 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 target languages: Artificial Neural Networks or Linear Discriminant Functions. The results obtained by each of these classifiers are then merged. The performance obtained with

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

this system is 12% EER on the NIST 2005 30-second excerpts.

The Lincoln Laboratory of the Massachusetts Institute of Technology has presented a fusion of several systems (Campbell et al., 2005). 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 (Gauvain et al., 2004)), 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-second 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 than language identification and has not been for the moment as much investigated (Lincoln et al., 1998; Berkling et al., 1998; Fung and Kat, 1999; Schultz et al., 2002; Wu et al., 2006; Zheng et al., 2006; Ikeno and Hansen, 2006; Huang and Hansen, 2006), although one can find a growing number of references on a related problem – foreign accent identification (Vieru-Dimulescu and de Mareüil, 2006). Many approaches use language identification systems applied to native dialect identification.

For example, in (Torres-Carrasquillo et al., 2004) 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 s utterances: American English (North vs. South), Chinese (Mandarin vs. Taiwan) and Spanish (Caribbean vs. Non-Caribbean). The error rate were respectively: 15.0% for American English, 11.5% for Chinese and 13.7% for Spanish.

In (Chen et al., 2001), 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 h. Sixty speakers were used for testing for each dialect. The results were between 12% and 15% errors (for female and male speakers, respectively) for utterances of approximately 20 s.

The experiments reported in (Tsai and Chang, 2002) concern also the identification of Chinese dialects. Here the considered dialects are Mandarin, Holo and Hakka (all spoken in Taiwan). The corpus used in these experiments is quite small, with a total of eight speakers reading 30 paragraphs, generating sentences about 15 s long. The same speakers are used for training and testing, and each speaker read each text three 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 Ph.D. thesis of the first author (Rouas, 2005a) was successfully tested on read speech from seven languages (English, French, German, Italian, Spanish from the original Multext corpus (Campione and Véronis, 1998), Mandarin Chinese (Komatsu et al., 2004) and Japanese (Kitazawa, 2002)), achieving around 70% of correct identification (Rouas, 2005b). The system was also tested on semi-spontaneous Arabic dialects (Araber database, Rouas et al., 2006), where the task was to discriminate between geographical areas linked to the dialects in three zones (Maghreb, Middle-East and Intermediate, i.e. Egypt and Tunisia), having achieved an area identification rate of 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.

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

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

- 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.

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 (Mateus and d'Andrade, 2000; Barbosa and Albano, 2004). 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* [sdʃprz′armuʃ] '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* [pisikoloʒ′iɐ] 'psychology', *afta* [′afitɐ] 'aphtha' in BP vs. [psikluʒ′iɐ], [′aftɐ] in EP). Loanwords can be treated rather differently, as well (e.g. [iʒn′ɔbi] in BP vs [sn′ɔb] 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 neutralized 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) [prefer′e~sjɐ] – *preferível (preferable)* [prifir′ivew]). Vowel height harmony in BP has been extensively studied, as it constitutes an important factor for the differentiation of BP dialects (Callou and Leite, 1990; Leite et al., 1996). 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, and speaker's age.). Vowel lowering is typical of northern dialects and is practically non-existent 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) [pe~s′ɐmuʃ]/ *pensámos* (we thought) [pe~s′amuʃ]);
(2) low vowels are raised before heterosyllabic nasal consonants (e.g. *cara* (face) [k′arɐ]/ *cana* (cane) [k"ɐnɐ]);
(3) front vowels centralize before palatal consonants and glides (e.g. *desenho* (drawing) [dz′ɐ ɟu], *telha* (tile) [t′ɐʎɐ], *lei* (law) [l′ɐj]).

In BP the two forms in (1) are homophones (*pensamos* [pe~sɐmus] or [pe~sɐmuʃ]), and provided the orthography

is correct the desired pronunciation is generated. In fact, although since (Lacerda and Rossi, 1958), it has often been pointed out that nasalization is much stronger in BP than in EP, it has also been shown (Abaurre and Pagotto, 1996) 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 [ɬ] 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 [ɐmiˈzadə] 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 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 [prĩspɐlmˈe~tə] in AP and as [prĩsipalm~ˈe~tə] 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.

3.3. Prosodic differences

The literature on the rhythm of Portuguese shows that there are controversial issues. In (Parkinson, 1988), for instance, EP is classified as stress-timed and BP as having mixed patterns of the syllable and stress-timed type. In (Frota and Vigário, 2001), 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 (Frota et al., 2002), 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

[2] F. Vicente, personal communication.

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 Fernandes, 2007), 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 (Rouas, 2007, 2006), we have decided to implement a prosodic system, conjointly with the PRLM and acoustic system.

The system is thus a fusion of three subsystems: Acoustic (Section 4.2), Phonotactic or PRLM (Section 4.3), and Prosodic (Section 4.4). These three subsystems use a common pre-processing module as represented in Fig. 1. The pre-processing module will be briefly reviewed in the following section:

4.1. Audio 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 audio pre-processing module also used in the speech recognition system. This module, developed by H. Meinedo (see Meinedo and Neto, 2003; Meinedo and Neto, 2005), integrates five components (Fig. 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 on a Portuguese BN corpus of over 60 h.

Two of the modules of this pre-processing stage are specially interesting for language identification: 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.

All the modules were evaluated by H. Meinedo on the COST 278 corpus described in Section 5. The acoustic change detector achieved a Recall value (% of detected acoustic change points) of 78.9%, a Precision value (% of detected points which are genuine change points) of 65.5%, and an F-measure (defined as $(2 * Recall * Precision)/(Recall + Precision)$) of 70.9. The speech/non-speech detector achieved an Accuracy of 95.6%, and the gender detector of 94.5%. Concerning the speaker clustering module, its performance was evaluated in terms of Q-measure (68.1%) and Diarization Error Rate (DER = 31.6%). 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 an incorrect cluster–speaker correspondence. As one speaker is often divided in several clusters,

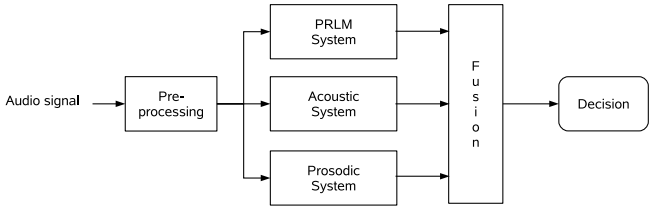

Fig. 1. Overview of the language identification system.

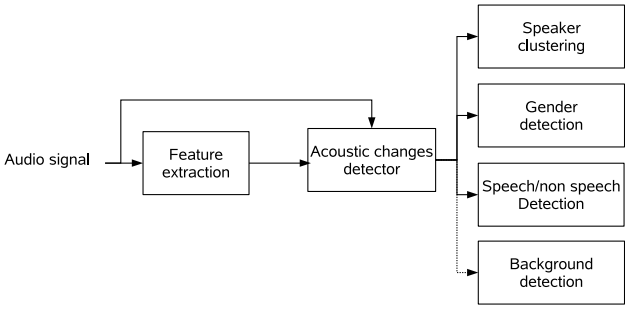

Fig. 2. Overview of the audio pre-processing module.

the performance in terms of DER is not very high, but this type of error does not affect the LID system.

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 (Meinedo and Neto, 2005).

4.2. Acoustic system

A generic acoustic language identification system is displayed on Fig. 3. The system works in two phases: a learning procedure to create the models, and a testing procedure. The acoustic features extracted from the audio signal are 12 MFCC plus delta, resulting in a 24-dimensional vector. The models used are Gaussian Mixture Models (as in Zissman, 1993), learnt with the classic VQ and EM algorithms.

The background model is learnt using excerpts from all languages, while the target model is learnt using only the target language. No adaptation is used is this case. The verification test is made by comparing the likelihood of the test excerpt to the target-language model and to the background model.

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

As explained above, the PRLM system is based on a single Portuguese phone-recognizer (see Zissman and Berkling, 2001 for a description of PRLM systems). A synoptic of the system is given in Fig. 4.

The phone recognizer is part of the AUDIMUS system (Meinedo et al., 2003). 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, trained on the above mentioned EP broadcast news corpus of over 60 h. It combines phone probabilities computed from several MLPs using different feature sets: PLP (12th order plus delta), log-RASTA (12th order plus delta), and Modulation Spectrogram (MSG – 28 coefficients). The outputs of all three MLP classifiers are then merged using an average in the log-likelihood domain.

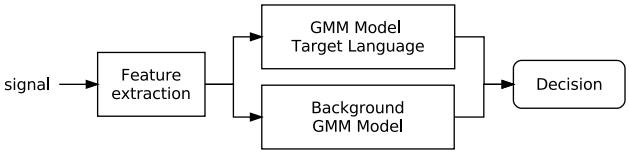

Fig. 3. Generic acoustic language verification system.

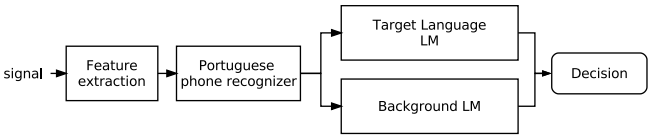

Fig. 4. PRLM system overview.

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 (Stolcke, 2002). A background *n*-gram model is also trained using data from all languages.

The verification test is made by comparing the likelihood of the test excerpt to the target-language model and to the background model. 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 consists of 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 (Rouas, 2007). 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 in Fig. 5.

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 (Dauer, 1983). Three baseline procedures lead to relevant consonant, vocalic and silence segment boundaries:

- Automatic speech segmentation based on the Forward–Backward Divergence" (DFB) algorithm (André-Obrecht, 1988), leading to infra-phonemic quasi-stationary segments.
- Vocal activity detection based on a first order statistic analysis of the energy signal (Pellegrino and André-Obrecht, 1997).
- Vowel localization based on a spectral analysis (Pellegrino and André-Obrecht, 1997).

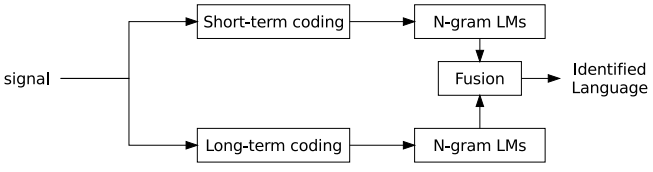

Fig. 5. Prosodic system overview.

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

This front-end processing results in a segmentation into vocalic, consonantal and silence segments. Labels "V", "C", or "#" are used to qualify each segment, respectively. The next stage is pseudo-syllable gathering: all the consonantal segments are merged until the next vocalic segment, which ends the pseudo-syllable. Fig. 6 shows the automatic segmentation and labeling results and the identified pseudo-syllables.

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 (Fujisaki, 2003). 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 (Sjölander, 2000).

The long-term coding uses the pseudo-syllable segmentation as a time-base. The coding is described in Fig. 7. The "baseline" is a representation of the phrase 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 in Fig. 6.

The short-term coding is detailed in Fig. 8. The short-term coding use the "C", "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. The F0 variation on voiced parts gives the label (Up or Down). Unvoiced parts are labeled "#". 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. The Up and Down labels are used to describe the variations while very short segments (e.g. <20 ms) are labeled "#". Duration labels are also computed on the segment units. The "s" (short) and "l" (long) labels are assigned considering the mean duration of each kind of segment (vocalic, consonantic or silence). These three coding labels are used conjointly to form the short-term coding. Hence, for each segment, the label is then composed of three symbols.

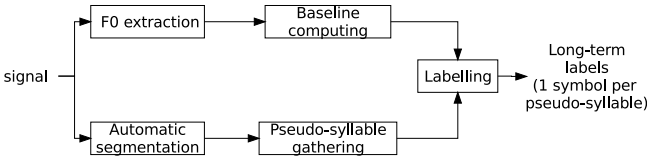

Fig. 7. Long-term coding.

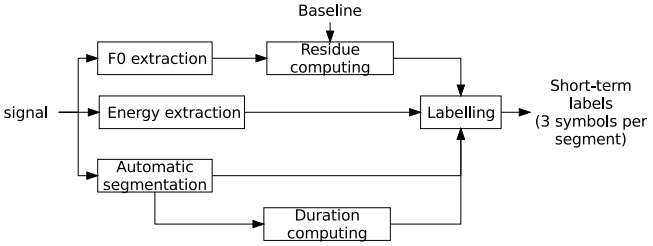

Fig. 8. Short-term coding.

To model the prosodic variations, we use classical *n*-gram language modeling provided by the SRI language modeling toolkit (Stolcke, 2002). For each system – long- and short-term – each target language is modeled by a *n*-gram model during the learning procedure. A background

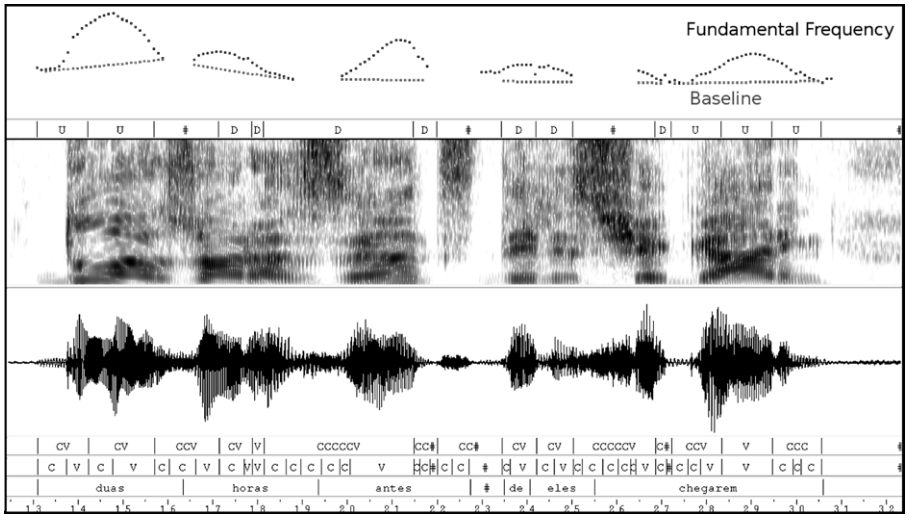

Fig. 6. Spectrogram, signal representation and prosodic processing example for the sentence "na mesma noite **duas horas antes de eles chegarem** uma casa havia sido assaltada na cidade". Transcriptions are (from bottom to top): (a) manual word annotation, (b) automatic segmentation and labeling, (c) pseudo-syllables, and on top of the spectrogram: long-term coding.

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

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. Several lengths for the *n*-gram models have been tested (from 3- to 5-gram), the best results are obtained with 3-gram on different kinds of databases (Rouas, 2007).

4.5. Fusion

For the time being, the fusion method is only a simple weighted addition of the log-likelihoods generated by each system. The weights have been computed on the train part of the corpus described in the next chapter. The method is clearly non-optimal. 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 6).

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.

5.1. Language verification corpora

The COST 278 corpus was constructed by seven institutions that collaborated in the European 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) (Vandecatseye et al., 2004). The first part of Table 1 shows the countries and languages used by the different TV stations, the number of collected shows and the total data size (in minutes).

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 subset of the 1996 Broadcast News Speech Corpus. This corpus contains a total of 104 h 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.

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.

[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 Republic	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	33 h 47 min

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.

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

The test set has a total duration of 7 h 15 min, with 789 automatically detected speakers. The duration per language ranges from 24 to 75 min.

5.3. Variety verification corpora

5.3.1. European Portuguese

For the variety verification task, we used the EP subset of the COST 278 corpus. Since this corpus includes different varieties of Portuguese, it was manually processed to eliminate the non-EP speakers. The duration is now 152 min, for the train set, and 21 min for the test set.

The short duration of the EP test set, relative to what became available for other varieties, led us to add an extra daily news show. The new EP corpus has a total of 230 min (78 for testing), and 336 automatically detected speakers.

5.3.2. Brazilian Portuguese

The Brazilian recordings come from news shows of the TV Record Brazilian channel. We have recorded 12 shows, ranging from 20 to 50 min each. After pre-processing, we have a total of 367 min of Brazilian speech data, with 452 automatically detected speakers.

5.3.3. African Portuguese

Reporter Africa is the main news programs from the RTP Africa channel. Each daily show lasts for around

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

Table 2
Portuguese varieties recordings from "Reporter África"

Code	Country	# Speakers (test)	Duration in minutes (test)
AN	Angola	84 (35)	71 (23)
CV	Cape Verde	116 (24)	78 (11)
GB	Guinea-Bissau	70 (26)	71 (22)
MO	Mozambique	77 (22)	77 (23)
ST	São Tomé and Príncipe	83 (31)	62 (18)
Total	5	430 (138)	359 (97)

25 min, with information from reporters in Angola, Cape Verde, Guinea-Bissau, Mozambique, São Tomé and Príncipe. The anchor speaks European Portuguese. We have recorded 24 shows and labeled the varieties for each of these. The total duration is 10 h, 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 2.

5.3.4. Train and test sets

Preliminary experiments with a reduced data set were first carried out using a cross-validation procedure (Rouas et al., 2008), given the relative low volume of data for each variety. First, one speaker was selected for testing. All the remaining data was used for learning the variety models. After the test was completed, a new speaker was used for testing. This procedure was iterated until all the speakers of the corpus have been used for testing. The problem with this first approach was that speakers were clustered independently for each show and we did not guarantee that the same speaker was not used both for training and testing.

The current data set is roughly 70% larger, which enabled us to have separate train and test sets. In selecting these sets, we tried to guarantee that the same speaker was not present in both train and test sets. The percentage of the corpus used for testing ranged from 14% to 34% for all the varieties.

6. Language verification experiments

The language verification results shown in this chapter were computed using the verification framework adopted in the recent NIST evaluation compaigns. Results include miss and false alarm probabilities, DET curves,[4] and equal error rates.

In the language verification corpus, 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 Fig. 9.

The best results are unsurprisingly given by the PRLM system, using 3-grams. The GMM-MFCC system gives the second best results, followed by the prosodic short-term system. The DET curve obtained using the fused system is also displayed in Fig. 9. Results are given in terms of EER in Table 3, for each system and each language. Different thresholds are used for computing the EER for each language.

The overall performance of the fused system is 12.4% EER. 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). Tests with much larger corpora should be made to evaluate if such differences in performance are significant.

Tables 4 and 5 show, respectively, the number of false alarms and missed detections over all languages and, given the focus of our work, for Portuguese. These results were obtained using the fused system.

The Portuguese false alarms are distributed across the different languages: one from English, nine from Galician, five from Greek, three from Sloven and one 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 s).

The only missed detection error for the Portuguese verification system appears on a 2-second segment, which is in

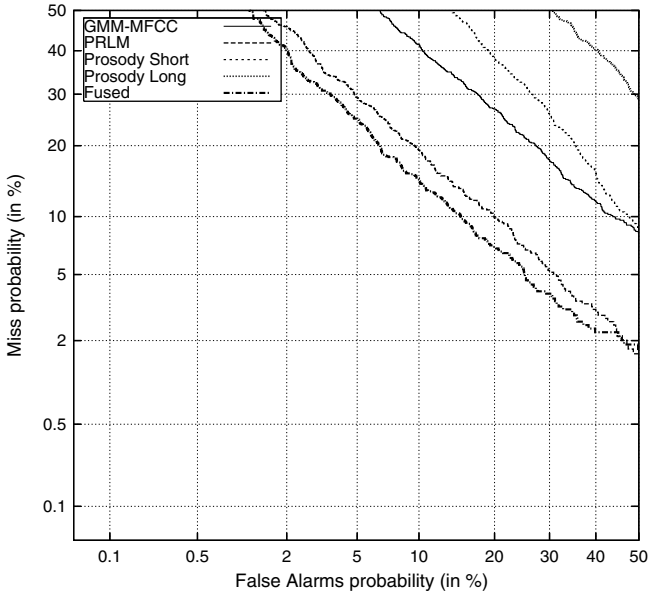

Fig. 9. DET curve obtained using all the systems.

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

Table 3
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

Table 4
False detections for the fused system

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

Table 5
Missed detections for the fused system

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

fact a music segment in English, wrongly labeled as speech by the pre-processing module.

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.

6.1. 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 10, 20 and 30 s.

The first line of Table 6 shows that the number of test segments reduces (for all languages) when only long segments are selected. As expected, the performance of the system shown in the second line improves when using longer segments. The improvement is clearly significant when selecting segments of duration over 30 s: the EER becomes 5.8% instead of 12.4%. This is illustrated by the DET-curve displayed in Fig. 10. Selecting segments that have enough information for identifying the language is clearly needed to achieve better performance.

Table 6
Results for all the systems in terms of % EER depending on the minimum duration of the test segments

Minimum duration	0 s	10 s	20 s	30 s
# Test segments (ALL)	789	618	394	272
ALL (% EER)	12.4	9.3	6.6	5.8
EP (% EER)	2.5	0.2	0.00	0.00

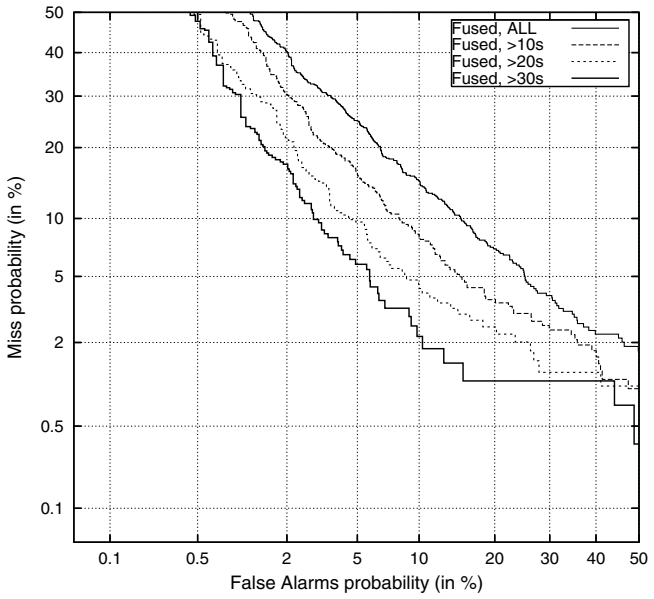

Fig. 10. DET curve obtained using the fused system.

The same type of analysis is shown in particular for the Portuguese language verifier in the last line of Table 6. One can observe that the system does not make any errors for files over 20 s, and that the error rate is only 0.2% for files over 10 s. 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.

7. 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 only used the Portuguese language verifier. The test data described in Section 5.3 is used for testing – thus, we have all African varieties speakers (138) plus Brazilian speakers (147) plus European Portuguese speakers (125). 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.

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

Table 7
Results for the system in terms of % EER depending on the minimum duration of the test segments

Minimum duration	0 s	10 s	20 s	30 s
# Test segments	410	228	152	94
Fused system (% EER)	20.7	16.6	13.8	10.6

Table 7 shows the results obtained for all the segments durations. When not selecting a minimum length, there is a total of 410 segments to test. The number of test segments becomes 228 when considering segments of duration superior to 10 s, 152 for segments superior to 20 s, and 94 for segments superior to 30 s.

The variety which is least recognized as Portuguese is Brazilian Portuguese, which 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.

8. Discrimination between Portuguese varieties

Given that the amount of data for each variety is not very well balanced (much more data for the European and Brazilian varieties than for each separate African variety), we first tried to address the problem of identifying three broad varieties by regrouping all the African varieties into one class. Thus, the aim of this experiment is to verify if the test speaker speaks African, Brazilian or European Portuguese. The designed system performs fairly well on this data, with a global identification rate of 83.9%. Detailed results (Table 8) show that the best identified variety is Brazilian Portuguese (96.6%). This result is obtained using the fused system.

African varieties tend to be confused with BP and, although no so often with EP. The next experiment aims at assessing the confusability between African varieties themselves. The global identification rate (see Table 9) is only 42.0%. The most clearly identified variety is the one from Cape Verde.

Table 8
Identification of Portuguese varieties – confusion matrix using only three broad classes (African, Brazilian and European Portuguese)

	AP	BP	EP
AP	79.7	12.3	8.0
BP	3.4	96.6	0.0
EP	14.4	12.0	73.6

% Correct = 83.9% ± 3.5%.

Table 9
Identification of African Portuguese varieties – Confusion matrix produced by the fused system

	AN	CV	GB	MO	ST
AN	31.4	20.0	5.7	20.0	22.9
CV	4.2	66.7	8.3	12.5	8.3
GB	3.8	42.3	30.8	15.4	7.7
MO	13.6	13.6	13.6	59.1	0.0
ST	9.7	29.0	3.2	25.8	32.3

8.1. 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 eight stimuli from each of the seven 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 s. Most of the sentences were extracted from spontaneous speech (64%), in order to avoid easily identifiable journalists or politicians. In addition to the eight sentences, the participants were asked to identify the variety of two words (also extracted from sentences). The total duration of all stimuli was 8.5 min. Participants were asked to classify each stimulus as one of the seven 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, currently living in Portugal. Forty-four participants were Portuguese, seven were from Brazil and fourteen from Africa (eight from Angola, four from Cape Verde and two from Mozambique).

Table 10 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.
- 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

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

Table 10
Human benchmark results (% of correct identification)

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

The top part shows the results with sentences and the bottom part shows the results with words.

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 11, 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 investigating 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%).

8.2. 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 in Section 4.2.

Table 11
Human benchmark results with only three broad classes

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

The top part shows the results with sentences (correct = 96.2%) and the bottom part shows the results with words (correct = 91.8%).

Table 12
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	42.8	73.5	43.0	19.8	42.7	40.6	44.3

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 (≈ 50 h). The language models were smoothed using Knesser–Ney discounting and entropy pruning. The perplexity obtained in a development set is 112.9.

Table 12 shows the ASR results in terms of word error rate (WER), obtained using all the training/test material of our accent identification 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 distinguishable from EP, both manually and automatically. Intermediate results were obtained for all African varieties, with very close WER values slightly above 40% for all of them. 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.

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

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

9. 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 30 s" 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 the automatic audio pre-processing system. When tested over segments of duration above 10 s, the equal error rate drops to 0.2% EER, and no errors were observed when considering segments above 20 s. 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 (10.6% 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 using only three broad classes achieved an average correct identification rate of 83.9%. The least confusable variety was by far BP (96.6% correct identification). EP was next. African varieties were the hardest to discriminate. When trying to discriminate between the African varieties themselves, the correct identification rate was only 42.0%.

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 mainly for BP, 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 (Gauvain et al., 2004) 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 back-end classifier using either Neural Networks or Fuzzy Logic algorithms.

Acknowledgements

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 also by the European project Vidi-Video, and by PRIME National Project TECNOVOZ No. 03/165.

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006

References

Abaurre, M.B., Pagotto, E.G., 1996. Nasalizaçào no Português do Brasil. In: Koch, I. (Ed.), . In: Gramática do Português Falado, Vol. VI. Editora da Unicamp/FAPESP, Campinas SP, pp. 495–526.

André-Obrecht, R., 1988. A new statistical approach for automatic speech segmentation. IEEE Trans. Acoust. Speech Signal Process. 36 (1), 29–40.

Barbosa, P., Albano, E., 2004. Brazilian Portuguese – illustrations of the IPA. J. Internat. Phonetic Assoc. 34 (2), 227–232.

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

Callou, D., Leite, Y., 1990. Iniciação à Fonética e Fonologia. In: Zahar, Jorge (Ed.), Rio de Janeiro.

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

Campione, E., Véronis, J., 1998. A multilingual prosodic database. In: Internat. Conf. on Spoken Language Process., Sidney. <http://www.lpl.univ-aix.fr/projects/multext>.

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

Dauer, R.M., 1983. Stress-timing and syllable-timing reanalysed. J. Phonetics 11, 51–62.

Fernandes, F., Subject localization constraints in BP and EP, Ph.D. thesis, Univ. Estadual Campinas.

Frota, S., Vigário, M., 2001. On the correlates of rhythmic distinctions: the European/Brazilian Portuguese case. Probus 13, 247–273.

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

Fujisaki, H., 2003. Prosody, information and modeling – with emphasis on tonal features of speech. In: ISCA Workshop on Spoken Language Process., Mumbai, India.

Fung, P., Kat, L., 1999. Fast accent identification and accented speech recognition. In: ICASSP'1999, Phoenix, AZ.

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

Huang, R., Hansen, J.H., 2006. Gaussian mixture selection and data selection for unsupervised spanish dialect classification. In: INTERSPEECH'2006, Pittsburgh, PA.

Ikeno, A., Hansen, J.H., 2006. The role of prosody in the perception of US native english accents. In: INTERSPEECH'2006, Pittsburgh, PA.

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

Komatsu, M., Arai, T., Sugawara, T., 2004. Perceptual discrimination of prosodic types. In: Speech Prosody, Nara, Japan.

Lacerda, A., Rossi, N., 1958. 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, pp. 5–102.

Leite, Y., Callou, D., Morais, J., 1996. Neutralização e realizaçã fonética: a harmonia vocálica no Português do Brasil. In: Actas do Congresso Internacional sobre o Português, Vol. III, Lisboa.

Li, J., Yaman, S., Lee, C.-H., Ma, B., Tong, R., Zhu, D., Li, H., 2006. Language recognition based on score distribution feature vectors and discriminative classifier fusion. In: Proc. Odyssey 06: The Speaker and Language Recognit. Workshop.

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

Matejka, P., Burget, L., Schwarz, P., Cernocký, J., 2006. Brno university of technology system for nist 2005 language recognition evaluation. In: Proc. Odyssey 2006: The Speaker and Language Recognit. Workshop.

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

Meinedo, H., Caseiro, D., Neto, J., Trancoso, I., 2003. Audimus.media: a broadcast news speech recognition system for the European Portuguese language. In: PROPOR'2003 – 6th Internat. Workshop on Comput. Process. of the Portuguese Language.

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

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

Parkinson, S., 1988. Portuguese. In: Harris, Martin, Vincent, Nigel (Eds.), The Romance Languages, pp. 131–169.

Pellegrino, F., André-Obrecht, R., 1997. Vocalic system modeling: a VQ approach. In: IEEE Digital Signal Process., Santorini.

Rouas, J.-L., 2005a. Caractérisation et identification automatique des langues. Ph.D. thesis, University Toulouse 3, France.

Rouas, J.-L., 2005b. Modeling long and short-term prosody for language identification. In: INTERSPEECH'2005, Lisboa, Portugal.

Rouas, J.-L., 2007. Automatic prosodic variations modelling for language and dialect discrimination. IEEE Trans. Audio, Speech Lang. Process. 15 (6), 1904–1911.

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

Rouas, J.-L., Trancoso, I., Viana, C., Abreu, M., 2008. Portuguese variety identificaiton on broadcast news. In: 2008 IEEE Internat. Conf. on Acoustics, Speech, and Signal Process. (ICASSP 2008), Las Vegas, Nevada.

Schultz, T., Jin, Q., Laskowski, K., Tribble, A., Waibel, A., 2002. Speaker, accent, and language identification using multilingual phone strings. In: HLT'2002, San Diego, CA.

Sjölander, K., 2000. The snack sound toolkit. URL <http://www.speech.kth.se/snack/>.

Stolcke, A., 2002. Srilm – an extensible language modeling toolkit. In: INTERSPEECH'2002, Denver, CO. URL <http://www.speech.sri.com/projects/srilm/>.

Torres-Carrasquillo, P.A., Gleason, T.P., Reynolds, D.A., 2004. Dialect identification using gaussian mixture models. In: Odyssey: The Speaker and Language Recognit. Workshop, Toledo, Spain.

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

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

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

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

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

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

Zissman, M.A., 1993. Automatic language identification using gaussian mixture and hidden markov models. In: IEEE 18th Internat. Conf. on Acoustics, Speech, and Signal Process., Minneapolis, MN, USA.

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

Please cite this article in press as: Rouas, J.-L. et al., Language and variety verification on broadcast news for Portuguese, Speech Comm. (2008), doi:10.1016/j.specom.2008.05.006