PORTUGUESE VARIETY IDENTIFICATION ON BROADCAST NEWS

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

[1] INESC-ID, Spoken Language Systems Laboratory (L2F), Portugal
[2] INRETS Electronic, Waves and Signal Processing Research Laboratory for Transport, France
[3] Centro de Linguística da Universidade de Lisboa, Portugal

ABSTRACT

This paper describes an accent verification system for Portuguese, that explores different type of properties: acoustic, phonotactic and prosodic. The system is designed to be used as a pre-processing module for the Portuguese Automatic Speech Recognition system developed at INESC-ID. 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. When distinguishing between European, Brazilian and African Portuguese, 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.

Index Terms— Language verification, Portuguese varieties.

1. INTRODUCTION

One of the problems encountered by the Automatic Speech Recognition (ASR) system developed at INESC-ID when applied to automatic captioning of broadcast news (BN) is the presence of different languages and different varieties of Portuguese. The presence of varieties other than European Portuguese (EP) may severely degrade the performance of the recognizer. In fact, whereas the word error rate (WER) of an ASR trained for EP is around 24% for this variety, for African Portuguese (AP) it can go from 30% to 38%, and for Brazilian Portuguese (BP), it may exceed 60%. This motivated the need for a variety identification module.

The orthographic differences are minor, which justifies similar out-of-vocabulary rates for the three varieties (1.4%, 2.0%, and 1.8%, for EP, AP and BP, respectively). Syntactic differences can be found in the use of prepositions, the position of clitics, and the alternative use of infinitive/gerundive verb forms. The lack of number agreement can be also found in BP and specially AP. However, the most striking differences concern pronunciation, namely vowel reduction, which is much more extreme in EP than in BP [1], [2]. Concerning prosody, whereas comparative studies of BP and EP can already be found [3], 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.

Dialect/accent identification is a somewhat harder topic that language identification and has not yet been as much investigated [4] [5] [6] [7] although one can find a growing number of references on a related problem - foreign accent identification. Many approaches use language identification (LID) systems applied to native dialect identification.

This was the approach that we also followed and that will be described in section 2. The next section is dedicated to the corpus used in our variety identification experiments. The results are discussed in section 4, and compared with a human benchmark test.

2. LANGUAGE IDENTIFICATION SYSTEM

Our LID system is a fusion of 3 subsystems: Acoustic (section 2.2), Phonotactic or PRLM (section 2.2), and Prosodic (section 2.4). These 3 subsystems share a common audio pre-processing module (APP) as represented in the Figure 1. For the time being, the fusion

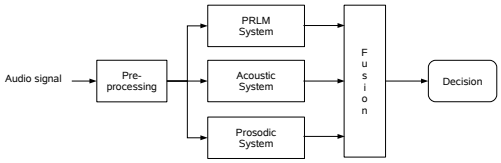

Fig. 1. Overview of the language identification system.

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.

2.1. Audio pre-processing

The APP module is part of our speech recognition system [8]). It integrates five components: three for classification (Speech/Non Speech, Gender and Background), one for speaker clustering and one for acoustic change detection. 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 hours [9]. Two of the modules of this pre-processing stage are specially interesting for LID: 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 variety and make the verification decision on a speaker by speaker basis.

2.2. Acoustic system

A generic acoustic language identification system is displayed on Figure 2. The system works in two phases: a learning procedure to create the models, and testing procedure. The acoustic features

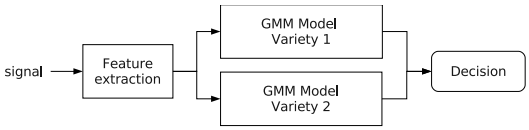

Fig. 2. Generic acoustic language verification system.

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 [10]), learnt with the classic VQ and EM algorithms.

2.3. PRLM system

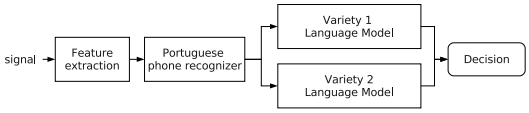

Fig. 3. PRLM System overview.

As explained above, the PRLM system is based on a single Portuguese phone-recognizer. A synoptic of the system is given in the Figure 3.

The phone recognizer is part of the AUDIMUS system [11], a hybrid recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification abilities of Multi-Layer Perceptrons. 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 [12].

2.4. Prosodic system

The prosodic system is the same as used in [13]. 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 4.

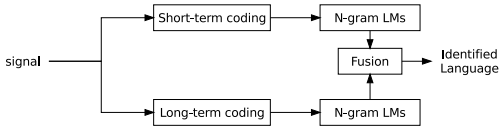

Fig. 4. Prosodic system overview.

2.4.1. 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 [14].

Three baseline procedures lead to relevant consonant, vocalic and silence segment boundaries: automatic speech segmentation [15], vocal activity detection [16] and vowel localisation (see [16] for more details). Labels "V", "C", or "#" are used to qualify each segment. Then, all the consonantal segments are merged until the next vocalic segment, which ends the pseudo-syllable.

2.4.2. 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 [17]. 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 [18].

The long-term coding uses the pseudo-syllable segmentation as a time-base. The coding is described in Figure 5. The "baseline" is a

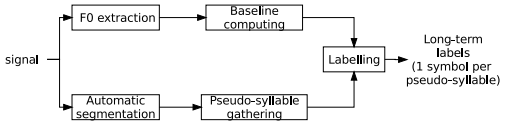

Fig. 5. Long-term coding.

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).

The short-term coding is detailed on Figure 6. The short-term

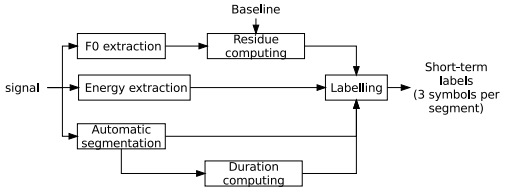

Fig. 6. Short-term coding.

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 labelled "#". 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. $<$20ms) are labelled "#". 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 are used conjointly to form the short-term coding. Hence, for each segment, the label is then composed of three symbols.

2.4.3. Prosodic N-gram Modeling

To model the prosodic variations, we use classical n-gram language modelling provided by the SRI language modeling toolkit [12]. As the best results are obtained with 3-grams on different kinds of databases [13], this setting has been kept for these experiments. For

each system – long- and short-term – each target variety is modeled by a n-gram model during the learning procedure. A background model is also learned using data from all varieties. During the test phase, the most likely variety is picked according to the model (target or background) which provides the maximum likelihood.

3. CORPORA

For the variety identification task, we used the EP subset of the COST 278 corpus [19], complemented with BN shows transmitted from Portugal to Brazilian (TV Record) and African (Reporter Africa) speakers. For EP, the corpus includes 6 shows of different type. For BP, the corpus includes 6 40-minute shows (including publicity). For AP, we have recorded 16 30-minute shows and labeled the varieties spoken by reporters in Angola, Cabo Verde, Guinea-Bissau, Mozambique, São Tomé and Príncipe. The number of automatically detected speakers and the duration (in minutes) for each variety is shown in Table 1.

Table 1. Portuguese varieties.

Code	Country	#spks	Dur(min)
EP	Portugal	296	192
BP	Brazil	404	190
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

4. VARIETY IDENTIFICATION EXPERIMENTS

The first experiment aimed 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: 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.

The global variety identification results are shown in Table 2. The average 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%).

Table 2. Identification of Portuguese varieties - Confusion matrix (%)

	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 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 next addressed 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 EP). Thus, the aim of this experiment is to identify if the test speaker speaks African, Brazilian or European Portuguese.

Table 3. 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 3) 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%.

After identifying that a speaker speaks African Portuguese, the third experiment aimed at finding which African Portuguese variety is actually spoken. The global identification rate (see Table 4) 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 4. Identification of African Portuguese varieties (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

4.1. 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 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. 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. 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 5 shows the confusion matrix.

Table 5. 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

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.

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. Hence, the bad results for EP identification. 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 6, the average ratio of correct identification is 96.2%.

Table 6. Human benchmark results with only 3 broad classes (correct=96.2%).

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

Just for comparison purposes, we have also run an experiment aimed at investigating the behavior of the automatic 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%).

5. CONCLUSIONS

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%.

6. REFERENCES

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

[2] P. Barbosa and E. Albano, "Brazilian Portuguese - illustrations of the IPA," *Journal of the International Phonetic Association*, vol. 34, no. 2, pp. 227–232, 2004.

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

[4] T. Wu et al., "Improving the discrimination between native accents when recorded over different channels," in *INTERSPEECH'2005, Lisbon, Portugal*, 2006.

[5] Y. Zheng et al., "Accent detection and speech recognition for shanghai-accented mandarin," in *INTERSPEECH'2005, Lisbon, Portugal*, 2006.

[6] A. Ikeno and J. H.L. Hansen, "The role of prosody in the perception of us native english accents," in *INTERSPEECH'2006, Pittsburgh, PA*, 2006.

[7] Rongqing H. and J. H.L. Hansen, "Gaussian mixture selection and data selection for unsupervised spanish dialect classification," in *INTERSPEECH'2006, Pittsburgh, PA*, 2006.

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

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

[10] M. A. Zissman, "Automatic language identification using gaussian mixture and hidden markov models," in *ICASSP*, Minneapolis, MN, USA, apr 1993.

[11] H. Meinedo et al., "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*, June 2003.

[12] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *INTERSPEECH'2002, Denver, Colorado*, 2002.

[13] J-L. Rouas, "Automatic prosodic variations modelling for language and dialect discrimination," *IEEE Trans. on ASLP*, vol. 15, no. 6, pp. 1904–1911, 2007.

[14] R. M. Dauer, "Stress-timing and syllable-timing reanalysed," *Journal of Phonetics*, vol. 11, pp. 51–62, 1983.

[15] R. André-Obrecht, "A new statistical approach for automatic speech segmentation," *IEEE Trans. on ASSP*, vol. 36, no. 1, pp. 29–40, 1988.

[16] F. Pellegrino and R. André-Obrecht, "Vocalic system modeling: A VQ approach," in *IEEE Digital Signal Processing*, Santorini, July 1997, pp. 427–430.

[17] H. Fujisaki, "Prosody, information and modeling - with emphasis on tonal features of speech," in *ISCA Workshop on Spoken Language Processing*, Mumbai, India, 2003, pp. 5–14.

[18] K. Sjölander, "The snack sound toolkit," .

[19] An Vandecatseye et al., "The COST 278 pan-european broadcast news database," in *LREC'2004*, Lisbon, 2004, pp. 873–876.