Exploiting variety-dependent Phones in Portuguese Variety Identification applied to Broadcast News Transcription

Oscar Koller[12], Alberto Abad[1], Isabel Trancoso[13], Céu Viana[4]

[1] L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
[2] Berlin University of Technology, Germany
[3] IST Lisboa, Portugal
[4] CLUL, Portugal

oscar@l2f.inesc-id.pt

Abstract

This paper presents a Variety IDentification (VID) approach and its application to broadcast news transcription for Portuguese. The phonotactic VID system, based on Phone Recognition and Language Modelling, focuses on a single tokenizer that combines distinctive knowledge about differences between the target varieties. This knowledge is introduced into a Multi-Layer Perceptron phone recognizer by training mono-phone models for two varieties as contrasting phone-like classes. Significant improvements in terms of identification rate were achieved compared to conventional single and fused phonotactic and acoustic systems. The VID system is used to select data to automatically train variety-specific acoustic models for broadcast news transcription. The impact of the selection is analyzed and variety-specific recognition is shown to improve results by up to 13% compared to a standard variety baseline.

Index Terms: accent identification, recognition of accented speech.

1. Introduction

Portuguese is the seventh most spoken language in the world [1]. But just about five percent of the Portuguese speakers live in Portugal and consequently speak Portuguese with an European accent. Automatic captioning of broadcast news (BN) faces heavy difficulties in the presence of different accents. The word error rate (WER) of L^{2}F's baseline European Portuguese (EP) recognizer degrades from under 20% with EP speech to around 30% for African Portuguese (AP) varieties, and above 50% for Brazilian Portuguese (BP). In order to overcome the challenges imposed by the presence of multiple varieties of Portuguese in BN data, variety dependent recognition systems and efficient variety identification modules are needed.

The PostPORT project (Porting Speech Technologies to other varieties of PORTuguese) focuses on these needs. At INESC-ID, we have been working for several years on Large Vocabulary Continuous Speech Recognition (LVCSR) using hybrid recognizers, combining Artificial Neural Networks and Hidden Markov models (ANN/HMM), the so-called connectionist paradigm. Our first LVCSR system was initially developed for European Portuguese, and recently ported to BP [2]. This paper reports on current advances in porting it to the AP varieties.

As manual transcriptions are expensive goods, a frequent alternative is unsupervised training of the acoustic models with automatically transcribed data, which has proven to have significantly lowered the WER in our EP recognizer. However, in the case of AP, we face the problem of only having access to broadcast news shows that include speakers from Portugal and five different African countries, with many regional differences within each national variety. The goal of this paper is to use a variety identification (VID) module to select stretches of speech representative of AP varieties that can be used for unsupervised training.

There are several approaches to tackle the problem of automatic language -or more specifically variety- identification that one can find in the literature. Most common approaches include acoustic, phonotactic or even prosodic based methods [3]. Phonotactic methods have been usually considered one of the best performing approaches.

Our VID approach is based on a combination of systems: a PRLM (Phone Recognition and Language Modeling) [4] system with a mono-phonetic tokenizer that was recently proposed in [5] and a Gaussian supervector based recognizer [6]. The VID module is used to select AP excerpts from untranscribed BN recordings. The selected parts are used together with a small manually transcribed corpus to train AP-specific acoustic models. The impact of the variety selection is evaluated and speech recognition results are compared to those obtained with the L^{2}F's baseline system trained for EP.

The motivation to consider AP as a broad geographical variety, instead of training a specific system for every African national variety is related to the reduced amount of training data and also to the fact that a human benchmark [7] revealed that identifying African varieties in BN is much harder than identifying accents of everyday's people on the street, possibly due to higher level of education and contact with EP in BN. Hence AP generally encompasses the varieties spoken in one of the five PALOP countries (African Countries with Portuguese as Official Language): Angola, Cape Verde, Guinea-Bissau, Mozambique and São Tomé and Príncipe. For the sake of space, this paper shall concentrate on the distinction between the two most confusable varieties, AP and EP, and ignore our previous work that also includes BP.

The following Section describes the corpora used for the task of variety identification, and for training the AP recognizer. In Section 3 our VID approach is briefly presented, and its application to AP/EP distinction is described, together with promising results. In the following section the mono-phonetic identification system is evaluated on a system using unsupervised training. Section 5 presents the main conclusions.

26 – 30 September 2010, Makuhari, Chiba, Japan

2. Corpora

2.1. Data for speech recognition

The AP BN corpus was recorded from RTP-África. The corpus consists of BN journals with focus on Africa. There are about 457 minutes of manually orthographically transcribed corpus for training, about 50 minutes for cross-validation and 48 minutes for testing. The testing corpus has been manually classified in noisy and clean conditions and in speakers having slight or strong accents. There are further 67 hours of BN recordings available, that contain mixed AP and EP speech and will serve for automatic unsupervised training.

Additionally, an equal amount of around 450 minutes of EP data has been selected to train the mono-phonetic phone recognizer described in next sections. The EP data is a sub-set of the corpus recorded in the scope of the European project ALERT that contains over 50h of BN material, recorded from RTP, the public TV channel in Portugal, all orthographically transcribed.

2.2. Data for variety identification

Tables 1 and 2 list the data used for training and testing the VID approach, respectively, in terms of total duration and number of segments. Care was taken to ensure that the same speakers do not appear in train and test simultaneously, and to avoid the inclusion of too many segments from the same speaker.

Train Data	AP	EP	$\sum$
duration [min.]	238.8	279.1	517.9
segments	1424	1283	2707
∅ dur./segm. [s]	10.1	13.1	11.5
<3s [%]	16.9	0.1	8.9
3-10s [%]	42.3	49.6	45.8
10-30s [%]	38.7	44.1	41.3
>30s [%]	2.2	6.3	4.1

Table 1: *Data to train the statistical AP and EP variety models.*

Test Data	AP	EP	$\sum$
duration [min.]	88.8	99.0	187.8
segments	610	412	1022
∅ dur./segm. [s]	8.7	14.4	11.0
<3s [%]	23.3	0.2	14.0
3-10s [%]	43.1	42.5	42.9
10-30s [%]	32.8	50.0	39.7
>30s [%]	1.0	7.5	36.2

Table 2: *Data for variety identification evaluation.*

3. Variety Identification

Our proposed system for variety identification combines a novel phonotactic PRLM approach with an acoustic approach based on Gaussian supervectors.

3.1. Mono-phonetic PRLM approach

Our PRLM approach focuses on the phonetic classifier, trying to build a system with a highly specialized tokenizer that incorporates the differences between AP and EP at this level. To better characterize these differences, we divide all occurring phones in our varieties into the following two groups [8]:

1. mono-phones: phones in one language/variety, that overlap little or not at all with those in another language/variety (e.g. the phonetic realization(s) of /r/ in English and German).
2. poly-phones: phones that are similar enough across the languages to be equated (e.g. [sh] in English and German)

Phonotactic approaches are able to benefit from both types of phones in order to classify speech. With the help of statistical modeling differently occurring poly-phones carry important information through the sequences they appear in. If mono-phones are found in speech they could, at least in theory, instantly help to differentiate our varieties.

Determining the set of phones, which is unique for a certain variety, given its neighboring varieties is not straightforward. Linguistic knowledge about the varieties' phonetic and phonological characteristics is crucial, but often not available, not sufficiently detailed or controversial. We use a computational method instead to find variety dependent unique phones. Binary multi-layer perceptrons (MLPs) are trained to discriminate between the same pairs of aligned phone classes. It is worth noting that the selected phones are not mono-phones from a strict linguistic point of view, but rather mono-phone-like units. For the sake of simplification the term 'mono-phones' is nevertheless used in the scope of this paper to refer to these units.

L^2F's baseline Automatic Speech Recognition (ASR) system [9] is used to align the training and development data of both varieties, although it is trained for EP. To train a binary classifier that allows us to see if an AP representation of a certain phone differs from its EP counterpart, we keep solely those two phones in our training data. All other phones are removed. The chosen two phones are given distinct output classes. After training, the successful separation of both classes can be verified using development data. This is a fast process, which enables the training of binary classifier to be performed for all 38 pairs of phones. In this way we can determine which phone classes are different enough to be successfully distinguished and hence contain mono-phonetic characteristics.

We choose the eight best performing phone classes as mono-phone units, namely [L], [O], [l], [e~], [J], [a], [e] and [Z] using Portuguese SAMPA (Speech Assessment Methods Phonetic Alphabet). After preliminary experiments with various numbers of mono-phones, this number seemed to be a good trade-off between network complexity, training data and classification performance. This leads to a phonetic recognizer with 30 poly-phones and two times eight mono-phones plus silence, thus 47 outputs, which seems reasonable considering the 14 hours (around seven per variety) of available training data. The mono-phonetic recognizer is the combination of three MLP outputs trained with Perceptual Linear Prediction (PLP),log-RelAtive SpecTrAl features (RASTA) and Modulation SpectroGram (MSG) features.

Phonotactics of each target language are modeled with a 3-gram back-off model, that is smoothened using Witten-Bell discounting.

3.1.1. Linguistic Interpretation of chosen Mono-Phones

Although it is possible to argue that, at an underlying phonological level, all Portuguese varieties share a common segmental inventory, some important differences may be found in the way those underlying segments are realized phonetically in different contexts. Some of these contextual variants are unique in the sense that they do not belong to the phone set common

to all varieties and, if correctly identified by language specific phone models, they may be used as important cues for accent identification.

In the experiments described above, [L] and [J] are among the best candidates to mono-phones, and effectively, /L/ and /J/ are frequently not pronounced as such by most AP speakers, but as a slightly palatalized lateral or nasal consonant followed by [j]. This pronunciation is identical to the one found when a /l/ is followed by a /i/, and that may partly explain why [l] also appears among the best candidates. Note, however, that a mono-phone for this consonant probably also accounts for lateral flaps. Otherwise, one could expect it would be ranked closer to other coronal consonants which are most often apico-alveolars in all AP varieties. This feature, particularly noticeable in /r/,/l/,/t/ and /d/ is often the only hint to the listener to identify AP speakers that otherwise do not differ from EP ones. It is surprising that /t/ and /d/ are not in a higher ranking.

Concerning vowel differences, besides the fact that in AP open-mid and close-mid vowels may present an intermediate quality, the constraints that regulate vowel contextual realizations are also different. Thus, for instance, while the presence of an /l/ in final syllable position blocks vowel reduction in EP, the same does not always happen in AP: e.g /a/ may be realized as /6/ or even /@/ in this context. Similarly, whereas in EP neither the secondary stressed vowels nor the linking vowel of morphological compounds may be reduced, in AP those compounds are often realized as single words and these vowels may surface either as mid or high (e.g rodoviária - [R%OdOvj"arj6] in EP, but [ruduvj"arj6] or [rodovj"arj6] in AP, depending on the speaker's linguistic profile).

3.2. Acoustic approach

Combining Gaussian mixture models (GMM) with Support Vector Machines (SVM) as a discriminative classifier [10], the so-called Gaussian supervector (GSV) approach, is a well-known state-of-the-art technique. In this approach, a Universal Background Model (UBM) is trained using large amounts of data from diverse sources, containing different languages, speakers and channels. For every single speech segment, the UBM is adapted using Maximum-a-Posteriori (MAP) criterion to the characterisitics of the utterance. Then, the adapted means and variances are stacked into a supervector form. Finally, the vectors coresponding to speech segments of the same language are trained against the vectors of all other languages using SVM with an appropriate kernel. The trained SVMs are used to perform the scoring of test super vectors extracted in the same way.

In this work, we have built a GSV system based on mean supervectors (MAP adaptation of the Gaussian means). Additionally, we have used an alternative scoring approach [6]. In constrast to the conventional GSV, each language SVM model is *pushed back* to a *positive* and a *negative* laguange-dependent GMM model, which are then used to calculate log-likelihood ratio scores. In certain situations, especially on short utterances, this approach (henceforth refered to as GMM-GSV) has shown improved accuracy.

3.3. Calibration and Fusion

Calibration and fusion of every language recognition system is performed with a single linear backend for all speech segments of different lengths. The FoCal Multiclass Toolkit [11] is used for this task. The final calibration and fusion weights correspond to the mean of five independent cross-fold calibrations, each using random 20% of the test set.

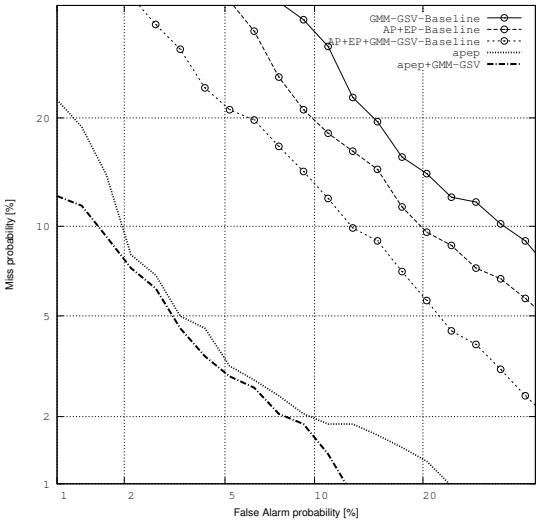

Figure 1: *DET-Curve for Identification of AP vs. EP*

3.4. Variety Identification Results

Fig. 1 shows the Detection Error Tradeoff (DET) curves for the identification of AP vs. EP, for several systems. The single mono-phonetic system is denoted as APEP. The single Gaussian supervector system is denoted as GMM-GSV. The fusion of two conventional phonotactic recognizers based on L^2F's EP phone tokenizer [9] and an AP phone classifier adapted from the EP system with the AP data presented in Section 2 is reported for comparison and it is denoted as AP+EP.

The fusion of the mono-phonetic system with the GMM-GSV outperforms all other approaches having an Equal Error Rate (EER) of 3.9%. However, this performance is mainly due to the single mono-phonetic classifier 'apep', which reaches an EER of 4.6%, being about 60% more successful than the baseline with 11.9% EER. Notice the great improvement achieved by the single mono-phonetic approach compared to the fusion of two individual phonotactic approaches with standard classifiers of both AP and EP varieties.

4. Application to Broadcast Transcription

The best performing approach for variety identification, which is the fusion of the mono-phonetic approach with the GMM-GSV, is used to select all AP parts from the untranscribed BN recordings presented in Section 2. The classifier determines 31 out of the total of 67 hours to be African accented speech. L^2F's baseline EP recognizer is used for automatic transcription of the selected data. The recognizer used a vocabulary of 100k words. The language model considered results from the interpolation of a 4-gram language model built from over 604M words of newspaper text, a 3-gram model based on manual BN transcriptions with 532k words, and a 3-gram based on about 560k words coming from automatic BN transcriptions.

Confidence measures based on a maximum entropy classifier are used to estimate the accuracy of each recognized word and to reject those with a confidence below a certain threshold. Nevertheless, the threshold applied produces 25% of wrong choices. As a result of this selection process, approximately half of the training data (17 hours of AP detected parts and 30 hours of the total corpus) remains to be used for training the

ASR system with automatic AP selection and without selection respectively. It is complemented by the limited available manually transcribed AP data of 457 minutes.

Between 1.2% and 4.6% improvement with VID compared to using no selection of the data is achieved, as shown in Tab. 3. Most significant improvements can be found particularly for slight accents and more noisy conditions. In our BN corpora, EP speakers often coincide with clean acoustic conditions. This fact probably accounts for higher improvement on noisy than on clean conditions by acoustic models trained on VID selected data, where there is much less speech in controlled environments than in non-selected recordings. Furthermore, the stronger impact of variety selection on slight accents is probably due to the fact that the VID system has been trained with slight and strong AP accents, where the former are much more present in our data. It has to be verified if focusing on strong accents in the identification process will further increase the impact of the VID system in this range.

The fact that AP and EP are very similar probably accounts for the moderate impact of VID in general. Moreover, systems with and without VID have been trained using the same manually transcribed data. However the influence of a bigger quantity and more diverse recordings, as well as a VID system tuned to stronger accents need to be investigated.

	any condition			clean		
accent	slight	strong	∅	slight	strong	∅
noVID	25.0	32.4	29.1	20.0	20.2	20.1
VID	23.9	31.8	28.1	19.6	20.0	19.7
change	4.6%	1.9%	3.2%	2.3%	1.2%	1.9%

Table 3: *WER Recognition results in [%] of AP-specific acoustic models with and without VID.*

Recognition results of L^{2}F's EP baseline and of the acoustic models trained using variety selection can be found in Tab. 4. Relative improvements range between around 6.4% on all conditions up to 13.2% on strong accents in clean conditions.

	any condition			clean		
accent	slight	strong	∅	slight	strong	∅
EP	25.4	34.0	30.1	20.5	23.0	21.5
AP	23.9	31.8	28.1	19.6	20.0	19.7
change	6.2%	6.5%	6.4%	4.4%	13.2%	8.2%

Table 4: *WER Recognition results in [%] of EP baseline and AP-specific acoustic models using variety identification.*

5. Conclusion

In this paper we presented an approach to exploit variety-dependent phones for the identification of AP and EP and used it for variety specific transcription of BN. We showed that a single mono-phonetic approach is able to reduce the EER to less than 60%, from 11.9% to 4.6%, of results achieved by our baseline, a fusion of standard phonotactic and acoustic approaches. Our approach further proves to be very efficient, employing just a single mono-phonetic tokenizer instead of three parallel systems.

Future work includes experiments with more data for train, calibration and test of the mono-phonetic system. Moreover a better understanding of how to choose the mono-phones is needed, as further improvement can possibly be achieved with different combinations and different number of chosen phones. We limited our number of mono-phones due to the amount of data available to train the phonetic classifier. Effects of more training data and consequently more mono-phones needs to be evaluated. Applying the mono-phonetic approach to varieties without available transcribed data, could be an interesting future investigation.

Applying the VID system to BN transcription proved to perform up to 4.6% better than a system using no data selection prior to training. It was shown, that the AP-specific recognition models lead to up to 13.2% relative WER improvement compared to using L^{2}F's EP baseline models for AP recognition.

It needs to be investigated if training the VID system solely with heavily accented speech can further strengthen its impact. Moreover the application without any manually transcribed data could lead to interesting results.

6. References

[1] M. P. Lewis, *Ethnologue: Languages of the World, 16th Edition*, 16th ed. SIL International, May 2009. [Online]. Available: http://www.ethnologue.com/

[2] A. Abad, I. Trancoso, N. Neto, and M. C. Viana, "Porting an European Portuguese broadcast news recognition system to Brazilian Portuguese," *Interspeech 2009, ISCA, Brighton, UK*, Sep. 2009.

[3] F. Castaldo, D. Colibro, S. Cumani, E. Dalmasso, P. Laface, and C. Vair, "Loquendo-Politecnico di Torino system for the 2009 NIST language recognition evaluation," *ICASSP 2010*, 2010.

[4] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 1, p. 31, 1996.

[5] O. Koller, A. Abad, and I. Trancoso, "Exploiting variety-dependent phones in Portuguese variety identification," in *Odyssey 2010: The Speaker and Language Recognition Workshop, 2010*, 2010.

[6] W. M. Campbell, "A covariance kernel for svm language recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, Mar. 2008, pp. 4141 –4144.

[7] J. Rouas, I. Trancoso, C. Viana, and M. Abreu, "Language and variety verification on broadcast news for Portuguese," *Speech Commun.*, vol. 50, no. 11-12, pp. 965–979, 2008.

[8] K. Berkling, T. Arai, and E. Barnard, "Analysis of Phoneme-Based features for language identification," *PROC ICASSP*, vol. 1, pp. 289—292, 1994.

[9] H. Meinedo, M. Viveiros, and J. Neto, "Evaluation of a live broadcast news subtitling system for Portuguese," in *Proc. Interspeech*, 2008.

[10] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, p. 210–229, 2006.

[11] N. Brümmer, "FoCal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition Scores—Tutorial and user manual—," 2007.