A NATIVENESS CLASSIFIER FOR TED TALKS

José Lopes[1 2], *Isabel Trancoso*[1 2], *Alberto Abad*[1]

[1]INESC-ID Lisboa
[2]Instituto Superior Técnico, Lisboa, Portugal
jose.david.lopes@l2f.inesc-id.pt

ABSTRACT

This paper presents a nativeness classifier for English. The detector was developed and tested with TED Talks collected from the web, where the major non-native cues are in terms of segmental aspects and prosody. The first experiments were made using only acoustic features, with Gaussian supervectors for training a classifier based on support vector machines. These experiments resulted in an equal error rate of 13.11%. The following experiments based on prosodic features alone did not yield good results. However, a fused system, combining acoustic and prosodic cues, achieved an equal error rate of 10.58%. A small human benchmark was conducted, showing an inter-rater agreement of 0.88. This value is also very close to the agreement value between humans and the best fused system.

Index Terms— Non-native accent, pronunciation.

1. INTRODUCTION

The goal of this paper is the automatic classification of speech as native or non-native. The study has been conducted for English, considering as native only American English speakers, and as non-native speakers from any other country which does not have English as the official language. Hence, for instance, speakers from UK and Australia, were not yet considered.

The study focused on studying non-nativeness in TED Talks. They are an ideal target for the current study, as they allow us to concentrate on particular aspects of non-native speech.

Non-native speech may deviate from native speech in terms of morphology, syntax and the lexicon, which is naturally more limited than for adult native speakers. In what concerns morphology, the main problems faced by non-native speakers concern producing correct forms of verbs (namely when irregular), nouns, adjectives, articles etc, especially when the morphological distinction hinges on subtle phonetic distinctions. The main difficulties in terms of syntax concern the structure of sentences, the ordering of constituents and their omission or insertion. In addition, non-native speech typically includes more disfluencies than native speech, and is characterized by a lower speech rate [1].

None of the above difficulties are very prominent in TED Talks given by non-native speakers, which are most often highly proficient in English. However, their speech frequently maintains a foreign accent, denoting the interference from the first language (L1), both in terms of prosody as well as segmental aspects. This justifies considering pronunciation one of the most difficult skills to learn in a second language (L2).

The segmental deviations in non-native speech can be quite mild, when speakers use phonemes from their L1 without compromising phonemic distinctions, but they may also have a strong impact on intelligibility whenever phonemic distinctions are blurred, a frequent occurrence when the phonetic features of L2 are not distinctive in L1.

Few studies target non-native accent identification using prosodic parameters [2, 3, 4, 5], for instance by automatically detecting erroneous word accent positions. Rhythmic traits, however, are generally regarded as the main source for the low intelligibility of L2 speakers. Although some authors base these difficulties on the differences in rhythm between L1 and L2, namely when one of the languages is stress-timed and the other is syllable-timed [6, 7], others claim more complex distinctions (for instance, syllables not carrying the word accent, that are weak in stress-timed languages, are produced stronger in syllable-timed languages).

Other non-prosodic approaches have been used in accent detection. Arslan and Hansen used word hidden Markov models (HMM) to identify accents [8], whereas Kumpf and King used HMM phone models to identify Australian English speech [9]. Bouslemi et al. tried another approach using discriminative phoneme sequences to detect speaker's origin through their accent when speaking English [10].

This work started with the development of a nativeness classifier based only on acoustic features, which was later extended to encompass prosodic features as well, borrowing from our previous experience in speaker and language recognition. Section 2 presents the corpus composition. Section 3.1 follows the steps used to build the acoustic classifier. Section 3.2 describes the prosodic features used and how they are incorporated in the classifier. In section 4, the results are shown and analyzed, and later compared with a human benchmark in section 5. Finally section 6 summarizes the main conclusions.

2. CORPUS

The collected corpus comprises 139 talks. The subset of native English speakers includes 71 talks, by US speakers. The remainder 68 talks are from non-native speakers. The fact that they may have lived in an English speaking country for some time was ignored in this classification, where place of birth was the major factor.

The first step in processing this corpus was to perform audio segmentation, separating speech segments from non-speech segments. Next, speech segments that were at least 1 second long were divided into train and test subsets, making sure that each talk was used only in one of the subsets.

More details on the corpus can be found in table 1, for the training and test sets, showing duration, total number of segments and percentage of segments for each of the durations considered.

Train	Native	Non-Native
Duration (min)	683.4	616.8
Segments (#)	1276	1299
<3 s (%)	2.1	4.8
3s-10s (%)	4.5	7.1
10s-30s (%)	60.0	64.9
>30 s (%)	33.3	23.3
Test	**Native**	**Non-Native**
Duration (min)	182.9	218.1
Segments (#)	400	548
<3 s (%)	4.5	2.2
3s-10s (%)	6.8	4.4
10s-30s (%)	62.2	74.1
>30 s (%)	26.5	19.3

Table 1. Details of the training and test sets for Native and Non-Native data.

3. NATIVENESS CLASSIFICATION

3.1. Acoustic Classifier Development

A method generally known as Gaussian supervectors (GSV) was first proposed for speaker recognition in [12]. In the last few years, however, this approach has also been successfully applied to language recognition tasks. GSV-based approaches map each speech utterance to a high-dimensional vector space. Support Vector Machines (SVMs) are generally used for classification of test vectors within this space. The mapping to the high-dimensional space is achieved by stacking all parameters (usually the means) of an adapted GMM in a single supervector by means of a Bayesian adaptation of a universal background model (GMM-UBM) to the characteristics of a given speech segment. In this work, we apply the GSV approach to the nativeness detection problem.

3.1.1. Feature Extraction

The extracted features are shifted delta cepstra (SDC) [13] of Perceptual Linear Prediction features with log-RelAtive SpecTrAl speech processing (PLP-RASTA). First, 7 PLP-RASTA static features are obtained and mean and variance normalization is applied on a per segment basis. Then, SDC features (with a 7-1-3-7 configuration) are computed, resulting in a feature vector of 56 components. Finally, low-energy frames (detected with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment) are removed.

3.1.2. Supervector extraction

In order to obtain the speech segment supervectors, a Universal Background Model (UBM) must be first trained. The UBM is a single GMM model that represents the distribution of speaker independent features. This is done in order to deal with the variability that characterizes accent recognition. All the training data available (native and non-native segments together) were used to train GMM-UBM of 64 and 256 mixtures, resulting in two different GSV-based systems.

One single iteration of Maximum a Posteriori (MAP) adaptation with relevance factor 16 is performed for each speech segment to obtain the high-dimensional vectors.

3.1.3. Nativeness modeling and scoring

Our classification problem is binary. Therefore we only need to train one single classifier. The linear SVM kernel of [12] based on the Kullback-Leibler (KL) divergence is used to train the target language models with the LibLinear implementation of the libSVM tool [14].

The native supervectors set is used to provide the positive examples, whereas the non-native set is used as background or negative samples. During test, the supervectors of the testing speech utterances are used by the binary classifier to generate a nativeness score.

3.2. Prosodic Classifier Development

Many of the previous approaches to nativeness classification that one can find in the literature use prosodic features. In [3, 4], these features are modeled with Hidden Markov Models. Lin and Wang [15] proposed a method for language identification aimed at modeling the prosodic contours (both energy and pitch) of syllable-like regions. In this work, we apply these contour information features for nativeness classification in order to complement the acoustic system described above.

3.2.1. Prosodic contour extraction

The Snack toolkit [16] is used to extract the log-pitch and the log-energy of the voiced speech regions of every utterance. Log-energy is normalized on an utterance basis. The prosodic contours are segmented into regions by splitting the voiced regions whenever the energy signal reaches a local minimum (the minimum length of the regions is 60 ms). We use a 3rd order derivative function of the log-energy to find local minima. For each region, the log-energy and log-pitch contours are approximated with a Legendre polynomial of order 5, resulting in 6 coefficients for each contour. The final feature vector is formed by the two contour coefficients and the length of the syllable-like region, which results in a total of 13 elements.

3.2.2. Nativeness modeling and scoring

Two different approaches were followed to train the nativeness detector that uses prosodic features. First, we trained GMM models for native and non-native speech models following the conventional GMM-UBM approach that is also applied in [15]. The UBM was estimated using all training data and the native and non-native GMMs were obtained based on MAP adaptation of the UBM with all the native and non-native training data, respectively. In this case, only a 64-mixture UBM was trained due to reduced amount of training vectors resulting from the fact that each feature vector now corresponds to a syllable-like segment of variable duration. The MAP adaptation was done with one iteration step. A second modeling approach was also developed based on the Gaussian supervector technique described in section 3.1.2 but now with the prosodic features. The MAP adaptation step for this approach uses the same UBM model of the previous approach.

During test, log-likelihood ratios of the native and non-native GMMs are computed for each testing speech utterance in the case of the GMM-UBM based system. In the GSV case, test supervectors are used with the binary classifier to generate a nativeness score.

3.3. Calibration

Each individual system was calibrated using the *s-cal* tool available with the Focal toolkit[17]. Additionally, Linear logistic regression

	Native		Non-Native		Total			
	GSV-acoustic 64	GSV-acoustic 256	GSV-acoustic 64	GSV-acoustic 256	GSV-acoustic 64		GSV-acoustic 256	
	Acc (%)	Acc (%)	Acc (%)	Acc (%)	Acc (%)	EER (%)	Acc (%)	EER (%)
<3s	35.7	35.7	80.0	80.0	47.4	24.3	47.4	30.0
3s-10s	66.7	48.2	66.7	81.0	66.7	28.3	62.5	30.7
10s-30s	82.7	88.4	87.0	91.4	85.3	15.1	90.2	10.1
>30s	89.6	91.5	85.9	82.1	87.7	10.4	86.8	7.6
Total	81.8	84.6	85.9	89.0	84.2	16.1	87.2	13.1

Table 2. Detailed results for GSV-acoustic 64 and GSV-acoustic 256.

is applied to the calibrated scores and it is also used for fusion of the acoustic and prosodic systems whenever it is necessary. A cross-validation strategy is carried out with the test data set to simultaneously estimate the calibration and fusion parameters and evaluate the system.

4. RESULTS AND DISCUSSION

This section describes the evaluation of the acoustic and prosodic systems for nativeness detection in TED talks, as well as the results of the fusion between them.

Table 2 presents experimental results targeted at comparing the performance of the acoustic-based Gaussian supervector system with different number of mixtures: GSV-acoustic 64 and GSV-acoustic 256. Both Accuracy (Acc) and Equal Error rate (EER) scores are computed for Native and Non-Native segments separately and for the overall test data. Results are also discriminated for different segment length durations.

These results show a generally better performance of the GSV-acoustic 256 system relative to the GSV-acoustic 64 classifier. In fact, increasing the supervector dimensionality allows considerable nativeness detection improvements. As expected, both classifiers perform better on longer duration segments, which is a well-known result of language recognition and other similar problems. Notice, however, that we have estimated duration-independent calibration parameters and that the performance loss in shorter segments could be partially alleviated if we had estimated length duration dependent calibrations. This effect can be particularly important given the fact that most of the test data segments have durations between 10 and 30 seconds. Finally, it is worth noticing that the detectors seem to be biased towards the non-native class for the shorter segments duration. This fact may be also related with a miss-calibration problem. In the longer segments duration case, both Native and Non-Native accuracies are quite balanced.

Table 3 shows the Accuracy and EER results obtained for the two prosodic systems: the Gaussian Supervector classifier (GSV-prosodic) and the GMM classifier (GMM-prosodic). In addition to the individual performance (column "alone"), the detection results of the prosodic based systems fused with the best acoustic system (GSV-acoustic 256) are also presented.

Results from Table 3 show that the performance of both prosodic systems alone is far below the one of the acoustic systems as it is clear also from DET curve in figure 1. However, when fused with GSV-acoustic 256, the combined system considerably improves the individual acoustic system.

Against our initial expectations, the GMM-prosodic performs better than the GSV-prosodic. This may be related to the reduced amount of prosodic feature vectors, which may influence the way MAP adaptation should be carried out for supervector extraction:

	alone		fusion	
	Acc.(%)	EER(%)	Acc.(%)	EER(%)
GSV-prosodic	68.9	40.7	89.1	10.6
GMM-prosodic	71.1	30.1	89.4	10.6

Table 3. Results obtained using prosodic features (Accuracy and EER) and the fusion between prosodic systems and GSV-acoustic 256.

number of iterations, relevance factor, etc. This possibility, together with the use of mean and variance normalized prosodic features, has been investigated but no conclusive results were obtained so far. Anyway, slight performance differences among the two prosodic systems are observed when fused to the acoustic classifier.

The final best performing nativeness detector is the one based on the fusion of the GSV-acoustic 256 with the GMM-prosodic subsystems as the DET curve in figure 1 is showing. The incorporation of the prosodic features permits an absolute performance improvement of 2.2% and 2.5% in terms of Accuracy and EER respectively, relative to the original detector based only on acoustic features.

These results are worse than the results for state-of-the-art language recognition, where EER is around 3% [18]. Recognizing non-native accents, however, is a far more difficult task.

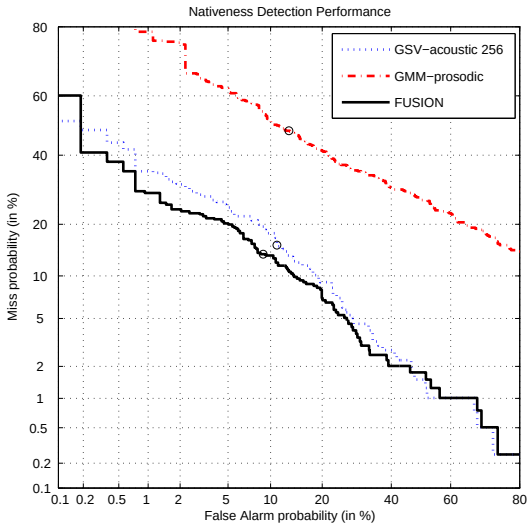

Fig. 1. DET curves of the GSV-acoustic 256, GMM-prosodic and fusion between both systems.

5. HUMAN BENCHMARK

It is interesting to compare the above results with the human performance on the same task. For that purpose, 3 native Portuguese speakers fluent in English, were asked to classify 45 segments. A set of 50 segments was randomly chosen from our test corpus, but 5 of them were discarded, because they contained semantic cues (e.g. one of the speakers actually said "I'm French").

The comparison of the original segment classification with the classification of each of the 3 subjects and the ones produced by the automatic classifiers produced discrepancies in only 10 files. Most differences occur in segments where the highly proficient speakers have a barely discernable non-native accent that also illudes some listeners. In this subset, the accuracy of the subjects averages 92.22%, whereas the one of the best fused system achieves 89.80%. We obtained a Cohen Kappa of 0.88 for the inter-rater agreement of the 3 subjects. The Cohen Kappa between the 3 subjects and the best fused classifier varied from 0.80 to 0.87.

6. CONCLUSIONS AND FUTURE WORK

This paper summarized our recent work on Native/Non-Native English classification. The first experiments were made using only acoustic features, with Gaussian supervectors for training a classifier based on support vector machines. These experiments resulted in an equal error rate of 13.11% which was later improved to 10.58% when prosodic features were also considered. Due to the reduced number of frames per file considered for computing the prosodic features it was difficult to discriminate between native and non-native segments using only this kind of feature.

The comparison between human performance and the developed systems shows that we are not far from human perception for this task.

The methods and the features were adapted from the ones typically used in speaker and language recognition problems. The results are lower than the ones obtained in language recognition, which may be due to the increased difficulty of the task of recognizing non-native accents, specially in a scenario where speakers are very fluent in English, many having lived for several years in a country that has English as a first language.

Expanding the study to encompass several native English accents (such as UK, Australian, etc.) is one of the next steps, despite rendering the problem even more difficult. The lack of data is one of the main difficulties in extending the study to other languages.

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues Miguel Bugalho and Helena Moniz. This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0053/2008 and CMU-Pt/0005/2007.

7. REFERENCES

[1] J. van Doremalen, C. Cucchiarini, H. Strik, "Optimizing automatic speech recognition for low-proficient non-native speakers.", EURASIP Journal on Audio, Speech, and Music Processing, 2010.

[2] F. Hoenig, A. Batliner, K. Weilhammer, and E. Noeth, "Automatic assessment of non-native prosody for english as l2", in Proc. Speech Prosody, Chicago, IL, 2010.

[3] M. Piat, D. Fohr, and I. Illina, "Foreign accent identification based on prosodic parameters", in Proc. Interspeech, Brisbane, 2008.

[4] J. Tepperman and S. Narayanan,"Better Nonnative Intonation Scores through Prosodic Theory", in Proc. Interspeech, Brisbane, 2008.

[5] F. Hoenig, A. Batliner, K. Weilhammer, and E. Noeth, "Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners", in Proc. SLATE, UK, 2009.

[6] F. Ramus, "Acoustic correlates of linguistic rhythm: Prespectives", in Proc. Speech Prosody, Aix-en-Provence, 2002.

[7] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis", in Laboratory of Phonoly VII, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton de Gruyter, 2002, pp. 515-546.

[8] L. Arslan and J. Hansen, "Language Accent Classification in American English", in Speech Communications, vol. 18, no-4, 1996, pp. 353-367.

[9] K. Kumpf and R. W. King, "Automatic Accent Classification of Foreign Accented Autralian English Speech", in Proc. ICSLP, 1996.

[10] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, "Discriminative Phoneme Sequences Extraction for Non-Native Speaker's Origin Classification", in ISSPA, 2007.

[11] D. A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing 10, 19-41, 2000.

[12] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification", IEEE Signal Processing Letters, Vol. 13, No. 15, May 2006.

[13] Torres-Carrasquillo, P. A. et alt., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", in Proceedings of ICSLP 2002, pp. 89-92, Denver, Colorado, Sep 2002.

[14] R.-E Fan, K.-W. Chang, C.-J Hsieh, X.-R Wang and C.-J Lin, "LIBLINEAR - A Library for Large Linear Classification", URL: http://www.csie.ntu.edu.tw/ cjlin/liblinear/.

[15] Lin C-Y, Wang H-C, "Language identification using pitch contour information", Proc. ICASSP 2005, Philadelphia, Pennsylvania, USA, 2005.

[16] Snack Toolkit v2.2.10, KTH Royal Institue of Technology, Department of Speech, Music and Hearing.

[17] Brummer, N., "Focal: Tools for Fusion and Calibration of automatic speaker detection systems",URL: http://www.dsp.sun.ac.za/ nbrummer/focal/.

[18] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition", Computer Speech and Language 20, 210-229, 2006.