Combining Multiple Approaches to Predict the Degree of Nativeness

Eugénio Ribeiro[1,2], *Jaime Ferreira*[1,2], *Julia Olcoz*[3],
Alberto Abad[1,2], *Helena Moniz*[1,4], *Fernando Batista*[1,5], *Isabel Trancoso*[1,2]

[1]L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]ViVoLAB, I3A, Universidad de Zaragoza, Spain
[4]FLUL/CLUL, Universidade de Lisboa, Portugal
[5]ISCTE-IUL – Instituto Universitário de Lisboa, Portugal
{eugenio.ribeiro, jaime.ferreira, alberto.abad}@l2f.inesc-id.pt

Abstract

Automatic speaker nativeness assessment has multiple applications, such as second language learning and IVR systems. In this paper we view this as a regression problem, since the available labels are on a continuous scale. Multiple approaches were applied, such as phonotactic models, i-vectors, and goodness of pronunciation, covering both segmental and suprasegmental features. Different phonotactic models were adopted, either trained with the challenge data, or using additional multilingual data from other domains. The obtained values were later combined in multiple ways and fed to a support vector machine regressor. Results on the test set surpass the provided baseline and are in line with the results obtained on the remaining sets. This suggests that our models generalize well to other datasets.

Index Terms: nativeness, phonotactics, GOP, prosody

1. Introduction

Knowing the degree of nativeness of a speaker is relevant for a number of applications. For example, a virtual language tutor could adapt selected materials according to the degree of nativeness of a student, making the lessons more compelling and productive. Since models are usually trained with native speech, it has a strong impact in Automatic Speech Recognition (ASR) tasks. Information about the degree of nativeness could be used by an ASR system to swap or adapt its language models, minimizing recognition errors in the presence of non-nativeness.

This paper reports our experiments in automatically identifying the degree of nativeness, in the context of the INTERSPEECH 2015 Computational Paralinguistics Challenge [1]. Since the challenge data is labeled in a continuous scale, we have tackled it as a regression problem. Multiple approaches were applied, such as phonotactic models, i-vectors, and Goodness of Pronunciation (GOP), covering both segmental and suprasegmental features. The resultant information was combined in multiple ways to feed a Support Vector Machine (SVM) regressor.

This paper is structured as follows: Section 2 presents the related work. Section 3 describes the datasets, features and approaches adopted. Results are presented and discussed in Section 4, and Section 5 concludes and presents future work.

2. Related Work

Nativeness assessment is a challenging task that has been explored for years. It is usually seen as a binary classification, predicting whether speakers are native. However, there has also been some work tackling this task in terms of the degree of nativeness, using both discrete and continuous scales. The latter is the focus of the challenge and, consequently, of this paper.

Literature on binary classification of nativeness uses distinct methods. Shriberg et al. [2] applied speaker recognition approaches to the non-nativeness detection task. The authors used systems based on Maximum Likelihood Linear Regression (MLLR) adaptation transforms, prosodic information, phone and word N-grams. By combining all systems, results around 8.6% Equal Error Rate (EER) were obtained. Omar and Pelecanos [3] were able to achieve 9.5% EER on the Fisher database, using an SVM classifier trained with ASR-based features integrated with an Universal Background Model (UBM). A sequential modeling of prosody to classify nativeness was proposed by Rosenberg [4]. In this approach, both symbolic and direct representations of prosody were used. Although symbolic representations outperformed direct ones, they were only able to do so when manual annotations were provided. This suggests that the first have limited scope, while the latter are more generically applicable, in spite of being less informative. Lopes et al. [5] developed a nativeness classifier for TED talks. Both prosodic features and Gaussian supervectors obtained from acoustic features were used. The fused system was able to achieve 10.6% EER. Mehrabani et al. [6], also using prosodic features, were able to exceed the accuracy of a Gaussian supervector by over 10.0%. The features were extracted from the Accent Group level, meaning, f0, energy, and duration were extracted from an accented syllable and all the following unaccented syllables until the next accent or boundary.

These former approaches were posed as binary classification tasks. However, we can also pose the degree of nativeness as a continuous task. Teixeira et al. [7, 8] assessed the importance of different prosodic features and their combinations for the task. Results were obtained using Decision Trees for both discrete and continuous scores. Of particular interest for the present work are the studies by Hönig et al [9, 10, 11]. The authors used the same data made available in this challenge, as well as acoustic-prosodic features. Therefore, their experiments are more directly comparable to the ones described in this paper. The first experiments applied Multiple Linear Regression on a large prosodic feature vector in order to automat-

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and under post-doc grant SFRH/PBD/95849/2013, EU-IST FP7 project SpeDial under contract 611396, and by the Spanish Government and FEDER through project TIN2011-28169-C05-02.

ically assess nativeness scores in terms of four dimensions – intelligibility, accent, melody, and rhythm. More recent experiments compare the performance of prosodic features, acoustic features extracted with openSMILE [12], and a Gaussian Mixture Model - Universal Background Model (GMM-UBM) trained with purely acoustic features to assess the levels of the same four dimensions. Further experiments were performed on the C-AuDiT and AUWL datasets assessing the influence of speaker and sentence dependency for the assessment of the rhythmic quality of non-native speech. In this sense, the evaluation performed in this challenge is both speaker and sentence independent. In that setup, using a cross-validation evaluation, the authors obtained 0.64 and 0.52 Person correlation coefficient scores on the C-AuDiT and AUWL corpora, respectively.

Summing up, nativeness has been analyzed either as a binary or a continuous prediction task, supported in distinct methods. However, the use of acoustic-prosodic features is almost transversal in the literature.

3. Experimental Setup

As the title of this paper suggests, we used different approaches in our experiments. These approaches have been used on this or similar tasks, but vary both in terms of nature and features used. Furthermore, we performed multiple combinations of the approaches in order to assess their complementarity. This section describes these approaches, the used datasets, and how the multiple feature sets were extracted.

3.1. Datasets

In addition to the provided datasets, some of our experiments involve the use of model-based features and approaches trained with external datasets. This is the case of acoustic models and phonotactic language recognition (LR) models that are part of the set of previously available technologies at the group.

3.1.1. Challenge datasets

The AUWL, ISLE, and C-AuDiT datasets are provided for the challenge and thoroughly described in the challenge's paper [1]. The first two are used for training while the last is used for development. The material from AUWL corresponds to 5.5 hours (3732 files) of pre-scripted dialogues spoken by learners of English as a second language. From ISLE, only 5 distinct sentences from 36 speakers are used, comprising 0.3 hours (158 files). C-AuDiT contains sentences read by non-native English speakers. 999 speech files containing 19 distinct sentences were selected. All the files were annotated by five phoneticians. However, while a five-point scale was used for the training data, a three-point scale was used for the development set.

3.1.2. Additional datasets

AUDIMUS The AUDIMUS acoustic modeling dataset consists of multilingual data used for training our in-house ASR system [13]. For the European Potuguese (*pt*) acoustic models, 57 hours of Broadcast News (BN) down-sampled data and 58 hours of mixed fixed-telephone and mobile-telephone data were used. The Brazilian Portuguese (*br*) models were trained with around 13 hours of BN down-sampled data. The European Spanish (*es*) networks used 36 hours of BN down-sampled data and 21 hours of fixed-telephone data. The American English (*en*) system was trained with the HUB-4 96 and HUB-4 97 down-sampled datasets, containing around 142 hours of data.

WSJ The Wall Street Journal (WSJ) database [14] is an US English native speakers database that contains high-fidelity speech recordings with excerpts from the Wall Street Journal. Only the SI-84 training material from WSJ0 was used for the development of a pronunciation quality measurement approach, consisting of approximately 15 hours of speech material.

euTV The euTV corpus consists of data used to develop the euTV [15] system for media monitoring and publishing. One of its services is able to identify the 12 most spoken languages across the European Union – English, Spanish, Polish, Greek, Portuguese, Hungarian, Czech, German, Italian, French, Dutch, and Swedish. Data was obtained from previously existing corpora used for automatic speech recognition, from the podcasts and archives made available online by the respective national radios and TV stations, and also from the podcasts and archives of the SBS [1] multi-language radio site.

LRE2011 The LRE2011 corpus consists of data used by INESC-ID's Spoken Language Systems Laboratory to develop the language recognition systems [16] submitted to the 2011 NIST Language Recognition Evaluation. It comprises data from 24 different languages obtained from different sources, including the data provided for the challenge; previous LRE campaigns; and several available Linguistic Data Consortium (LDC) sets.

3.2. Features

Our experiments use both segmental and suprasegmental features extracted from each speech file.

OpenSMILE [12] was used with the ComParE 2013 configuration file to extract the features also used by the existing baseline approach. The HTKToolkit [17] was used to extract Mel-Frequency Cepstral Coefficientss (MFCCs), and a module from our hybrid ASR system AUDIMUS [13] was used to extract Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) features. For each feature set, 7 static values were extracted. After that, mean normalization was applied in a per file basis. Shifted Delta Coefficients (SDCs) were computed from each feature set using a 7-1-3-7 configuration, originating two new feature sets. Finally, low-energy frames were detected with the alignment generated by a bi-Gaussian model of the log energy distribution computed for each speech file, and then discarded.

We performed a multi-language phone tokenization using the neural networks that are part of AUDIMUS [13]. The recognizer for each language combines four Multilayer Percepton (MLP) outputs trained with Perceptual Linear Predictions (PLPs) (13 static + 1st deltas), RASTA-PLPs (13 static + 1st deltas), Modulation-Filtered Spectrogram (MSG) (28 static), and ETSI (13 static + 1st and 2nd deltas). A phone-loop grammar with phoneme minimum duration of three frames is used for phonetic decoding. The language-dependent MLP networks were trained with the AUDIMUS dataset described previously. Each MLP network is characterized by the size of its input layer that depends on the particular parametrization and the frame context size (13 for PLP, RASTA-PLP and ETSI; 15 for MSG), the number of units of the two hidden layers (500), and the size of the output layer. MLPs are composed by two hidden-layers with a relatively small number of hidden units in order to accelerate the tokenization process. In this case, only monophone units are modelled, resulting in MLP networks of 41 (39 phonemes + 1 silence + 1 respiration) soft-max outputs in the case of *en*, 39 for *pt* (38 phonemes + 1 silence), 40 for *br* (39

[1] http://www.sbs.com.au/

phonemes + 1 silence) and 30 for *es* (29 phonemes + 1 silence).

The obtained phonetic tokenizations provide phone alignments for each speech file which can be used to generate more complex features. Based on the tokenizations, we have created what we have called *pseudo-words* and that consist on grouping phones into sequences separated by silences for each one of the languages (*en*, *pt*, *br*, *es*). From the phone alignments we were also able to extract a set of durational features. For instance we were able to extract silence and speech duration ratios, namely silence ratio, speech ratio, and silence to speech ratio. Furthermore, we were able to extract the speech rate in terms of phones per second, either taking or not pseudo-words into account. The phone tokenizations also provided means to characterize each audio segment using n-grams of phones for each one of the languages. As elongations may play a relevant role in nativeness detection, we have created additional n-gram variations whenever phones longer or equal than a given threshold (200 ms) were involved. For example, when in the presence of the phone *n* followed by the elongated phone *ay*, the bigrams *n.ay* and *n.ay+* would be activated.

Finally, we also used the Snack Sound Toolkit[2] as an alternate way of extracting pitch and energy from the speech signal. We have calculated pitch and energy metrics, such as maximum, minimum, standard deviation, range, and slope, inside and between pseudo-words [18]. Pitch related results were calculated based on semitones rather than frequency. On top of such features, we calculated elaborated prosodic features for the whole sentence, involving the sequence of calculated pseudo-words, that were measured in terms of standard deviation and slope.

3.3. Approaches

Although we used multiple approaches in our experiments, the final labels are always given by a SVM regressor. This means that the approaches described in this section were used to produce features to be used by that regressor. Furthermore, it is important to notice that some of the features described in the previous section are very informative by themselves and were used directly, without influence of any of these approaches. Finally, it is also important to refer that, for each experiment, we explored different combinations of the regressor's C (complexity) and epsilon parameters.

3.3.1. Phonotactic Models

To our knowledge, phonotactic models have not been used for degree of nativeness assessment. However, they have good performance on language recognition tasks and, thus, we believe they can be used for this task. Our models are obtained using Phone Recognition followed by Language Modeling (PRLM) [19]. To develop such systems, we exploit the information provided by the phonetic tokenizers described in Section 3.2. First, phonetic sequences are obtained for every file. Then, for each target language and for each tokenizer a different phonotactic n-gram language model is trained with the training sequences.

In order to train the models, we split the training data in classes relative to the level of nativeness. We used both a 2-class – Native and Non-Native – and a 3-class – Good, Average, and Bad – splits. Using these splits, we were able to train a model for each class and for each tokenizer language. The outputs of these models for the development and test datasets are the likelihoods of a given speech segment fitting each model. These likelihoods can be used as features for the SVM regressor.

Furthermore, we used euTV and LRE2011 models for each of the present languages, which output the likelihood of a given speech segment belonging to that language. These likelihoods can also be used as features for the regressor.

3.3.2. Goodness of Pronunciation

The GOP algorithm is widely used in Computer-Assisted Language Learning (CALL) systems [20] for automatic assessment of phone pronunciation comparing speakers' realizations with native phone models [21, 22]. In this work, GOP-based features have also been considered for the Degree of Nativeness task. In order to estimate them, we set up a four-step process. First, the conventional GOP algorithm is used to obtain phone-level confidence measures, by aligning reference phones from manual transcriptions and free-loop recognized phones. Second, we conduct a phone-dependent mean score normalization. The phone means are estimated taking into account all the phone realizations of the training corpus. Then, GOP values are first grouped at word-level, considering three different arithmetic operators: sum, average, and maximum. These three operators create three word-level GOP-based features. Finally, three sentence-level features are obtained for each arithmetic operator by simply accumulating the GOP values of all the words in a given sentence. This approach exploits monophone acoustic models for American English trained with the WSJ corpus using the HTKToolkit [23].

3.3.3. Phone N-gram Sequences

In this approach each audio segment was described in terms of using n-grams of phones, ranging from unigrams to trigrams, based on the previously produced phone tokenizations for each language. Elongated phones with duration above 200ms were represented using additional n-grams, as described in Section 3.2. The n-gram counts (or simply their binary presence) were then used as features for the SVM regressor, discarding the ones not present in the training set. However, this process led to more than 80 thousand features, considering all possible languages (*en*, *pt*, *br*, *es*), making experiments very slow. For that reason, most of our experiments restrict n-grams to those occurring at least 10 or 25 times in the training data.

3.3.4. I-Vectors

In the experiments using this approach, we used both SDC sets as features. The first step of this approach consists on training a GMM-UBM using all the training data. We performed experiments with different numbers of mixtures, ranging from 64 to 1024. Next, zero and first-order sufficient statistics are computed from the training data and used to estimate the T matrix. In order to do so, 10 Expectation-Maximization (EM) iterations are applied. In the first 7 iterations only Maximum Likelihood (ML) estimation updates are applied, while in the last 3 EM iterations both ML and minimum divergence updates are applied. The covariance matrix is not updated in any of the EM iterations. We also performed experiments with different total variability sub-space dimensions, ranging from 16 to 400. The estimated T matrix is then used for extraction of the total variability factors of all speech data. Finally, the resulting factor vectors are normalized to be of unit length. These vectors are referred to as i-vectors. The i-vectors extracted from the provided datasets were later used as features for the regressor.

[2] http://www.speech.kth.se/snack/

	development		test
	PCC	ρ	ρ
OpenSMILE ComPaRe 2013	0.403	0.415	0.425
Phone-based speech rates (SR)	0.570	0.565	
[P1] SR, pwords, sil.ratio, energy	0.591	0.609	0.559
[P2] SR, pwords, sil.ratio, energy, F0	0.588	0.597	0.557
Phone N-gram Sequences	0.406	0.457	
Phonotactic 2-Class	0.379	0.435	
Phonotactic 3-Class	0.394	0.443	
Phonotactic euTV	0.543	0.589	
Phonotactic LRE2011	0.543	0.589	
Phonotactic euTV + LRE2011	0.544	0.589	
Goodness of Pronunciation (GOP)	0.305	0.368	
Normalized GOP (NGOP)	0.260	0.277	
i-Vectors: RASTA-PLP SDC	0.145	0.135	
i-Vectors: MFCC SDC	0.214	0.222	
SR + euTV + LRE2011	0.576	0.609	
SR + euTV + LRE2011 + GOP	0.580	0.621	**0.580**
SR + euTV + LRE2011 + NGOP	0.599	**0.638**	0.564
[P1] + euTV + LRE2011 + NGOP	0.605	**0.638**	**0.576**
[P2] + euTV + LRE2011 + GOP	0.617	**0.644**	0.559

Table 1: Results obtained by the most relevant approaches.

4. Results

We performed multiple experiments combining the different features and approaches described in the previous sections. Table 1 summarizes the results achieved by the most relevant approaches on the development and test sets. The Spearman's Correlation Coefficient (ρ) is our primary evaluation measure [1], but the Pearson Correlation Coefficient (PCC) is also presented for the development set. The first line shows results for the baseline approach. The first relevant point to notice is the importance of prosodic features, shown on the second group of approaches. Both speech rate and silence ratio were able to surpass the baseline on their own. However, our Speech Rate experiments worked better with multiple language alignments produced by our phone tokenizers, while Silence Ratio performed better when only the Portuguese alignment was used. A possible explanation is that the Portuguese tokenizer is the one with more effort put into and the most accurate one, which makes its silence detection more reliable.

All phonetic-based approaches, shown on the third group of approaches, were also able to surpass the baseline. N-gram phone sequences produced acceptable results, but required several thousands of features, which increased considerably the regressor's training time when compared to other approaches. For that reason, reported experiments restricted N-gram features to those occurring at least 25 times in the training data. Such an approach relies on sequences seen during the train, which may constitute a problem when generalizing to other data. The phonotactic models trained using the provided train dataset performed well in cross-validation (ρ>0.70). However, such value reduced considerably when the evaluation was performed on the development dataset, suggesting that the models needed to be trained with more data in order to generalize well. In terms of the class split, the system using 3 classes performed slightly better than the 2-class one. This means that the introduced entropy is beneficial. Furthermore, it is important to notice that performance increased with each new language model added, which suggests that using data provided by phone tokenizers for other languages would improve the overall performance. Further evidence of this are the results obtained by the models trained using euTV and LRE2011 data, which contain larger training datasets and different languages. Also, by combining the models trained with those datasets, a more robust system can be build, improving generalization capabilities.

Goodness of Pronunciation approaches were not able to surpass the baseline on their own, which can be surprising since pronunciation is a very important factor to identify non-native speech. The most surprising factor is the negative effect of normalization, for which we have no plausible explanation.

Although i-vector approaches have performed well in language identification tasks, the results obtained in this task have been disappointing. Although we performed experiments using multiple combinations of the number of Gaussian mixtures used to train the UBM and the number of total variability subspace dimensions, the results were always in line with the ones presented in Table 1, which are far from the baseline.

By fusing some of the previous approaches we were able to further improve the results. For instance, merging the combined euTV and LRE2011 phonotactic model with the speech rate, we were able to achieve ρ>0.60. By appending non-normalized GOP features to that system, that result improved to 0.621. Surprisingly, appending the normalized GOP features, which had worse performance on their own, improved that result to 0.638. This represents a 54% relative improvement over the baseline.

In order to have more training data, we merged the train and development datasets by linearly scaling the development set labels. Using this approach, our best trial on the test set achieved ρ=0.580, representing a 36% relative improvement over the baseline. Such result was achieved by merging the speech rate, euTV and LRE2011 phonotactic models, and unnormalized GOP. Contrarily to what happened on the development set, using normalized GOP decreased the score to 0.564.

Finally, it is important to compare our results with related work. Our 3-class phonotactic models trained with the provided data achieved a 0.75 PCC score when evaluated using cross-validation. Under the same evaluation conditions, Hönig et al [11] obtained a 0.52 PCC score. This represents a relative improvement of 44%. By performing cross-validation on the development dataset, we were able to achieve a 0.66 PCC score. Although not completely comparable, this result is still slightly better than the 0.64 score obtained by Hönig et al [11] on the C-AuDiT corpus.

5. Conclusions

This paper reports experiments towards the advance of the state of the art on the degree of nativeness assessment task. This was partially achieved using approaches that, to our knowledge, had not been used for this task. In this sense, although phonotactic models have been proven efficient, especially when merged with other approaches, large amounts of training data are needed to develop models that generalize well to different datasets. Furthermore, we were able to confirm the importance of prosodic features, such as speech rate, which achieved high results on its own.

In terms of results, we surpassed the baseline for the development set by over 0.2, which represents a relative improvement over 50%. Also, our best trial on the test set surpassed the baseline by 0.17, representing a relative improvement of 36%.

We applied multiple approaches and different features. However, more have been left unexplored which could improve our results. For instance, we have not explored the potential of Gaussian supervectors, which have been proven to perform well in some related work. In terms of features, we have not explored extraction at the Accent Group level, which has also been proved efficient. We leave these and other possible approaches for future work.

6. References

[1] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition," in *Proceedings INTERSPEECH 2015, ISCA, Dresden, Germany*, 2015.

[2] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Abkbacak, "Detecting nonnative speech using speaker recognition approaches," in *Proc. of IEEE Odyssey-08 Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, 2008.

[3] M. K. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," in *ICASSP*. IEEE, 2010, pp. 4398–4401.

[4] A. Rosenberg, "Symbolic and direct sequential modeling of prosody for classification of speaking-style and nativeness," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 1065–1068.

[5] J. Lopes, I. Trancoso, and A. Abad, "A nativeness classifier for TED talks," in *ICASSP*. IEEE, 2011, pp. 5672–5675.

[6] M. Mehrabani, J. Tepperman, and E. Nava, "Nativeness classification with suprasegmental features on the accent group level," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13*, 2012, pp. 2073–2076.

[7] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sönmez, "Prosodic features for automatic text-independent evaluation of nativeness for language learners," in *ICSLP*, 2000.

[8] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sonmez, "Evaluation of speaker's degree of nativeness using text-independent prosodic features," in *Proc. of the Workshop on Multilingual Speech and Language Processing*, 2001.

[9] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic assessment of non-native prosody for English as L2," in *Proc. of Speech Prosody*, 2010.

[10] F. Hönig, T. Bocklet, K. Riedhammer, A. Batliner, and E. Nöth, "The automatic assessment of non-native prosody: Combining classical prosodic analysis with acoustic modelling," in *INTERSPEECH*. ISCA, 2012.

[11] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic assessment of non-native prosody - annotation, modelling and evaluation," in *Proc. of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT), Stockhold*, 2012, pp. 21–30.

[12] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838.

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

[14] D. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *DARPA Speech and Natural Language Workshop*, 1992.

[15] M. Bertini, A. D. Bimbo, G. Ioannidis, E. Bijk, I. Trancoso, and H. Meinedo, "euTV: a system for media monitoring and publishing," in *ACM Multimedia*, A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D. A. Shamma, M. Worring, and R. Zimmermann, Eds. ACM, 2013, pp. 453–454.

[16] A. Abad, "The L2F language recognition system for NIST LRE 2011," in *The 2011 NIST Language Recognition evaluation (LRE11) Workshop, Atlanta, US*, 2011.

[17] S. Young, "The htk hidden markov model toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[18] F. Batista, H. Moniz, I. Trancoso, and N. J. Mamede, "Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts," *IEEE Transactions on Audio, Speech and Language Processing, Special Issue on New Frontiers in Rich Transcription*, vol. 20, no. 2, pp. 474–485, feb. 2012.

[19] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[20] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in English pronunciation," in *Proc. of International Conference on Spoken Language Processing*, 1990.

[21] S. Witt, "Use of Speech Recognition in Computer-Assisted Language Learning," Ph.D. dissertation, University of Cambridge, Dept. of Engineeringambridge, 1999.

[22] S. Witt and S. Young, "Phone-level Pronuciation Scoring and Assessment for Interactive Language Learning," *Speech Communication*, vol. 30, pp. 95–118, 2000.

[23] S. Young, "The HTK hidden markov model toolkit: Design and philosophy," Cambridge University, Department of Engineering, Tech. Rep., 1993.