Discriminating Native from Non-Native Speech Using Fusion of Visual Cues

Christos Georgakis
Dept. of Computing
Imperial College London
cg2212@imperial.ac.uk

Stavros Petridis
Dept. of Computing
Imperial College London
sp104@imperial.ac.uk

Maja Pantic
Dept. of Computing / EEMCS
Imperial College London /
Univ. Twente
London, UK / Enschede, NL
m.pantic@imperial.ac.uk

ABSTRACT

The task of classifying accent, as belonging to a native language speaker or a foreign language speaker, has been so far addressed by means of the audio modality only. However, features extracted from the visual modality have been successfully used to extend or substitute audio-only approaches developed for speech or language recognition. This paper presents a fully automated approach to discriminating native from non-native speech in English, based exclusively on visual appearance features from speech. Long Short-Term Memory Neural Networks (LSTMs) are employed to model accent-related speech dynamics and yield accent-class predictions. Subject-independent experiments are conducted on speech episodes captured by mobile phones from the challenging MOBIO Database. We establish a text-dependent scenario, using only those recordings in which all subjects read the same paragraph. Our results show that decision-level fusion of networks trained with complementary appearance descriptors consistently leads to performance improvement over single-feature systems, with the highest gain in accuracy reaching 7.3%. The best feature combinations achieve classification accuracy of 75%, rendering the proposed method a useful accent classification tool in cases of missing or noisy audio stream.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition—*Applications*; J.m [**Computer Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Non-Native Speech; Visual-only Accent Classification; Foreign Accent Detection; Visual Speech Processing.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2655026.

1. INTRODUCTION

Accent can be identified through a set of pronunciation, articulation, intonation, lexical stress and rhythmic patterns that are common in the speaking style of individuals belonging to a particular language group. Identifying accented speech has emerged as a need to overcome limitations posed by mispronunciations on the efficacy of speech recognisers [4]. Beyond serving as a pre-processing step for speech recognition, accent analysis is essential for applications such as computer-assisted second language learning [14].

Most related work has viewed accent identification as a multiclass classification problem that aims to classify a speech sample to either the native accent of the target language or to one of separately modelled foreign accents. Those approaches mainly use Hidden Markov Models trained on acoustic features, such as prosodic and cepstral features [1]. More recent works [13, 10] have borrowed inspiration from language and speaker recognition to target binary discrimination between native and non-native speech. Shriberg *et al.* [13] employ maximum likelihood regression and phone N-gram features. Omar and Pelecanos [10] use a novel universal background model with Support Vector Machines to detect non-native speakers and their native language.

All the above research on accent classification has ignored features derived from the visual stream. However, the beneficial role of visual information to speech comprehension has been well documented and experimentally validated [12]. Specifically, automated visual-only approaches have been developed for language identification [8]. Another study shows that visual identification of accent is feasible for human observers [5]. These findings indicate that there are visual accent-sensitive cues that can be used to distinguish between native and non-native speakers.

In this paper, subject-independent accent classification experiments are conducted on continuous reading speech samples from the MOBIO Database [7], all captured by mobile phones. Static appearance descriptors are extracted and fed into Long Short-Term Memory Neural Networks (LSTMs) [3] for classification. Our results show that accuracy increases significantly, when accent predictions rely on majority voting from networks that have been separately trained with different appearance features.

2. OVERVIEW OF OUR METHOD

We present a fully-automated solution for discrimination between native and non-native speech in English, using decision-level fusion of visual features. The proposed system is graph-

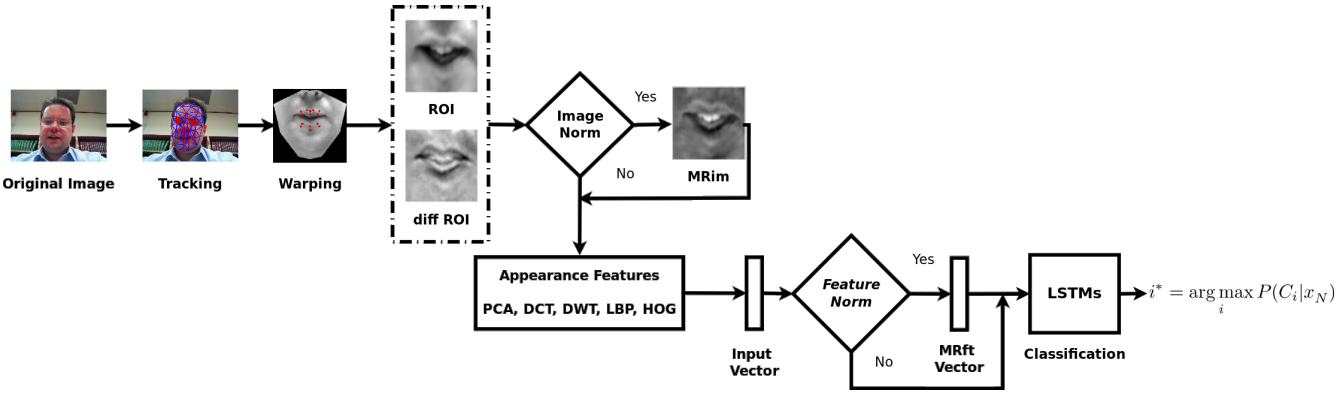


Figure 1: Illustration of the proposed framework for visual-only discrimination between native and non-native speech. The dash-dot line denotes a stage that involves variants, i.e., original or *difference* ROIs.

ically illustrated in Fig. 1. First, a texture warping process yields frontal face images, and the pixel intensities around the lips are used as mouth *Region of Interest* (henceforth termed mouth ROI). In the next stage, we use the mouth ROIs to extract five different appearance features. Different feature normalisation variants are also examined for the appearance descriptors. Finally, static features are forwarded to LSTMs, which assign a single accent label to each sample.

3. DATABASE

The proposed framework is evaluated on speech episodes in English from the MOBIO Database [7], by means of subject-independent experiments. This bimodal database was recorded at six sites in two phases, each comprised of six sessions. For each of the 150 participants, there are totally 192 recordings in English, almost exclusively captured on mobile phones. The visual stream is characterised by high variability in pose and illumination across frames, due to the acquisition device being handheld, while the appearance of subjects and background vary across sessions.

In the current study, we choose to include only the visual speech samples from Phase I in which all subjects read the same text, thus establishing a text-dependent experimental scenario. The data used are balanced over the two classes, with 135 samples belonging to 28 native English speakers and 137 to 28 non-native English speakers. The mean and standard deviation of duration over all samples is 22.5 and 3.4 seconds, respectively. Each video, encoded in variable framerate of mean value 15 fps, is converted in a sequence of still frames. All 272 samples used correspond to utterances of the following paragraph: *"I have signed the MOBIO consent form and I understand that my biometric data is being captured for a database that might be made publicly available for research purposes. I understand that I am solely responsible for the content of my states and my behaviour. I will ensure that when answering a question I do not provide any personal information in response to any question."*

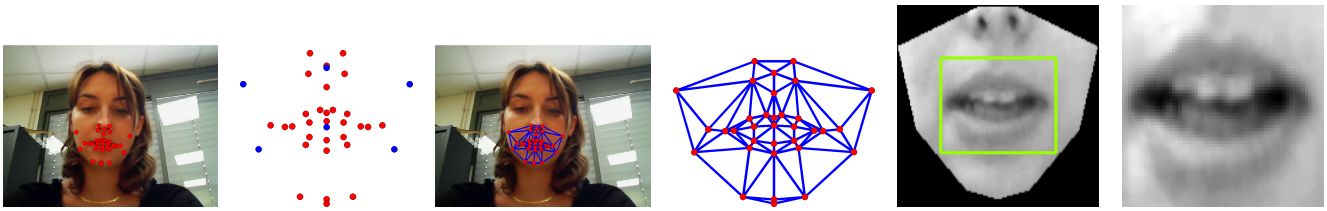
4. STATIC FEATURE EXTRACTION

We initially track 113 characteristic facial points, using the Appearance-Based Tracker [11]. These are manually annotated in the first frame and tracked for the remaining frames. We only use 34 points that correspond to the lower face region, specifically their 2D spatial coordinates (Fig. 2a), along with the coordinates of their pose-free version (Fig. 2b), all provided as a part of the tracker's output. All 34 pose-free points are globally registered according to the location of six base points (see blue points in Fig. 2b), which are relatively invariant to facial deformations.

Next, texture warping is performed to acquire lower face images in frontal view. First, for each frame, two 2D meshes (one for actually tracked points and one for aligned pose-free points), are triangulated. A piecewise affine warp is defined between the corresponding triangles. This warp is then used to map the texture of the mesh in the input image (Fig. 2c), onto the pose-free mesh (Fig. 2d). Each warped lower face is re-sampled to dimension 200×200, and the mouth ROI is extracted as a 94×114 bounding box containing the pixel intensities around the mouth (Fig. 2e). Finally, all mouth ROIs are downsampled to dimension 64×64 (Fig. 2f).

Five different appearance descriptors, all calculated on pixel intensities of the mouth ROIs, are examined; Principal Component Analysis [12], 2D Discrete Cosine Transform [12], Discrete Wavelet Transform [12], Local Binary Patterns [9], and Histograms of Oriented Gradients [2]. For *PCA*, we use those principal components accounting for the 95% of the total intensity variance. For *DCT*, the 2D cosine transform is applied to 8 non-overlapping 32×16 blocks of the ROI. Only four coefficients corresponding to the lowest frequencies are retained for each block and concatenated into a 32-dimensional vector. For *DWT*, ROIs are first rescaled to dimension 16×16, and a Daubechies-4 filter, with 3 levels of decomposition, is used. Our 64-dimensional vector consists of the approximation coefficients of the 3rd level and all the detailed coefficients of the 2nd and 3rd level. For *LBP*, we use the $LBP^{u2}_{(8,1)}$ scheme, which acts in a neighbourhood of 8 pixels on a circle of 1-pixel radius [9]. A 59-bin histogram, encoding the frequency of occurence of the $u2$-"uniform" LBP patterns over the entire ROI, forms our local texture descriptor. For the computation of *HOG*, four orientation bins are used and each ROI is divided into four 32×32 cells, so that dimensionality is comparable to that of the other features (64-element vector).

In order to capture the dynamics of the recorded visual articulations (i.e. changes in the skin appearance around the mouth), even prior to the classification stage, and remove redundant speaker-specific information, we also use *difference ROIs* [6]. These are computed for all frames as the difference in pixel intensities between the current ROI and the



(a) (b) (c) (d) (e) (f)

Figure 2: Instances of the ROI extraction process, illustrated on a speech frame from a non-native speaker of MOBIO Database. (a) Actually tracked points, (b) Pose-free points (the 6 base points used for alignment are shown in blue), (c) Triangulated mesh on the lower face image, (d) Triangulated mesh of the aligned pose-free points, (e) Warped frontal lower face and mouth bounding box, (f) Final rescaled mouth ROI.

Table 1: Distribution of Native and Non-Native subjects/samples over the three sets.

	Native	Non-Native	All
Training	14/71	14/65	28/136
Validation	8/32	7/36	15/68
Test	6/32	7/36	13/68
All	**28/135**	**28/137**	**56/272**

ROI at the previous frame. Mean normalisation [12] is also evaluated. The mean feature vector over all speech frames is subtracted from the feature vector of each frame. We call this scheme *mean removal at the feature level (MRft)*. We also examine the alternative of *mean removal at the image level (MRim)* [6], where the mean intensity over the entire utterance is removed from each ROI. Therefore, four different variants for each appearance feature are examined (*MRft, MRim, diffMRft, diffMRim*), with *diff* denoting the use of *difference ROIs*.

5. CLASSIFICATION WITH LSTMS

Long Short-Term Memory Neural Networks (LSTMs) [3] are in principle a variant of traditional recurrent neural networks. They have shown increased ability in capturing contextual statistical regularities in speech, even when those manifest themselves in longer time lags. Their hidden layers contain memory blocks, which are in turn composed of memory cells and three multiplicative gates that perform the operations of write, read and reset.

In the experiments reported in this paper, LSTMs are trained using the RNNLIB Toolbox [3]. The input layer has the size of the input feature vectors, which are corrupted with Gaussian noise of standard deviation 0.6 to improve generalisation. The sigmoid function σ is used for the activation α of the output layer unit, essentially transforming it into the posterior probability $P(C_1|x_n) = \sigma(\alpha)$ for the first accent class (native), where x_n is the feature vector at the n-th frame. Each speech example is classified as either native or non-native according to the value of this score for the vector of the last frame.

In particular, we use networks with one hidden layer. Weights are randomly initialised three times in the range $[-0.1, 0.1]$, and training is done with online gradient descent, with learning rate 10^{-4} and momentum 0.9. The number of blocks in the hidden layer is optimised on the Validation Set in the interval $\{40, 50, \ldots, 180\}$, separately for each feature-normalisation combination. The optimal number of blocks is set to the value yielding the lowest average classification error over the three trials on the Validation Set.

6. EXPERIMENTS

We evaluate the proposed framework by means of subject-independent experiments on visual speech data from MOBIO Database (see section 3). We use exactly 50% of the samples for training. The remaining episodes are equally divided into the Validation Set and the Test Set. All three sets are balanced, in terms of both accent class and gender of the subjects. The distribution of subjects and samples for each class across sets is shown in Table 1.

First, the Validation Set is used to find the optimal LSTM configuration, i.e., the number of blocks in the hidden layer, for each feature-normalisation combination and, subsequently, to reveal the best-performing normalisation for each feature. Results on the Validation Set, in terms of classification accuracy, are presented in Table 2. It is worth noting that more complex networks are needed to model feature vectors that vary less smoothly over time, such as the LBP and HOG local descriptors calculated on the *difference* ROIs (the optimal number of blocks for both $\text{LBP}_{diffMRft}$ and $\text{HOG}_{diffMRft}$ is 180). Furthermore, the *MRim* scheme seems to be more beneficial for features that are more susceptible to registration artefacts and misalignments, such as LBP and PCA. Instead, HOG is not assisted by further image-level processing of the ROIs, since its computation has already catered for intra-ROI illumination normalisation.

The normalisations that account for the highest accuracy percentages (shown in boldface in Table 2), are used for the corresponding features for the Test Set. Results on the Test Set, as yielded by networks trained on a single feature as well as all three- and five-element decision-level combinations, are reported in Table 3. Note that the F1 measure takes similar values for the two classes, indicating that the trained networks are not biased towards one of the classes. LSTMs trained either with DCT or with HOG achieve the highest accuracy of 67.7% amongst the single-feature systems. DCT is well-known for its ability to efficiently encode visual speech dynamics [12]. On the other hand, HOG proves also highly discriminative, presumably thanks to accent-related edge information being captured, such as tale-telling transient features (e.g., bulges and wrinkles). PCA accounts for the lowest accuracy of 60.3%, probably due to its higher susceptibility to registration errors.

The fusion results are obtained by means of majority voting, that is, each test example is assigned to the accent-label predicted by the majority of the three or five networks. As can be seen in Table 3, fusion consistently results in higher accuracy, compared to the single-feature systems. The only exception is the PCA+DCT+DWT combination, which still stays at the same accuracy as that obtained by

Table 2: Results in terms of classification accuracy (%) on the Validation Set of MOBIO speech samples, for the different features and normalisations examined. The accuracy reported for each combination corresponds to the network with the optimal number of memory blocks in the hidden layer, which is shown in the subscript. The best score for each feature among all four normalisations is shown in boldface.

Features/ Normalisations	*MRft*	*MRim*	*diffMRft*	*diffMRim*
PCA	$64.2_{(100)}$	$\mathbf{68.1}_{(70)}$	$65.2_{(100)}$	$66.2_{(70)}$
DCT	$\mathbf{65.2}_{(40)}$	$65.2_{(120)}$	$56.9_{(50)}$	$56.9_{(80)}$
DWT	$67.2_{(70)}$	$67.2_{(130)}$	$\mathbf{71.6}_{(100)}$	$71.6_{(140)}$
LBP	$59.8_{(60)}$	$\mathbf{64.7}_{(60)}$	$61.7_{(180)}$	$60.3_{(60)}$
HOG	$\mathbf{71.6}_{(120)}$	$56.9_{(70)}$	$60.3_{(180)}$	$59.3_{(60)}$

the best-scoring out of the three networks (67.7% by DCT). Fusion leads to higher performance in terms of the other metrics as well (class-wise F1 measures and UAR), in almost all cases. The largest performance boost, in terms of accuracy, over the single-feature schemes, amounts to 7.3%, and is furnished by both the PCA+DCT+LBP and PCA+DCT+HOG combinations. This highlights that those features capture complementary accent-sensitive information in the visual stream. Thus, a projection onto prominent modes of intensity variance (PCA), a block-based frequency decomposition (DCT), and a local descriptor unveiling local texture (LBP) or edge orientation information (HOG), train LSTMs that act efficiently in synergy for the target problem. The five-network combination does not provide additional gain, probably due to redundant information being carried by the couples (DCT, DWT) and (LBP, HOG).

7. CONCLUSION

In this paper, we presented a visual-only approach to discriminating native from non-native speech in English, based on fusion of neural networks trained on visual features. Overall, our results on fixed-content speech episodes provide evidence that accent in speech can be accurately identified when sequential classifiers, trained on complementary appearance descriptors, are combined to yield predictions. We intend to examine how our visual-only framework compares to an identical audio counterpart and experiment with multimodal fusion schemes. We also plan to investigate alternative ways of sequential modelling.

8. ACKNOWLEDGMENTS

This work has been funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA). The work of S. Petridis is also supported in part by EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching.

9. REFERENCES

[1] L. M. Arslan and J. H. Hansen. Language accent classification in American English. *Speech Com.*, 1996.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 2005.

[3] A. Graves. RNNLIB: A recurrent neural network library for sequence learning problems. `http://sourceforge.net/projects/rnnl/`.

Table 3: Results on the Test Set of MOBIO speech samples, reported as percentages: Class-wise F1 measure (F1-N for Native, F1-NN for Non-Native), Unweighted Average Recall (UAR), and Classification Accuracy (Acc.), for the various single-feature LSTMs as well as fusion of three and five of them. The highest value achieved for each metric is shown in boldface.

Features	F1-N	F1-NN	UAR	Acc.
PCA	69.7	42.6	62.3	60.3
DCT	67.7	67.7	67.9	67.7
DWT	62.3	69.3	65.8	66.2
LBP	62.3	69.3	65.8	66.2
HOG	65.6	69.4	67.5	67.7
PCA + DCT + DWT	70.3	64.5	68.4	67.7
PCA + DCT + LBP	77.3	72.1	75.9	**75.0**
PCA + DCT + HOG	**77.9**	71.2	**76.0**	**75.0**
PCA + DWT + LBP	70.3	64.5	68.4	67.7
PCA + DWT + HOG	71.2	66.7	69.8	69.1
PCA + LBP + HOG	72.5	71.6	72.4	72.1
DCT + DWT + LBP	70.8	73.2	72.1	72.1
DCT + DWT + HOG	71.0	**75.7**	73.3	73.5
DCT + LBP + HOG	71.6	72.5	72.2	72.1
DWT + LBP + HOG	67.7	73.0	70.3	70.6
All	76.1	73.9	75.5	**75.0**

[4] V. Gupta and P. Mermelstein. Effects of speaker accent on the performance of a speaker-independent, isolated-word recognizer. *J. Acoust. Soc. Am.*, 1982.

[5] A. Irwin, S. Thomas, and M. Pilling. Identification of language and accent through visual speech. In *Speech Prosody*, 2007.

[6] P. J. Lucey. *Lipreading across multiple views*. PhD thesis, 2007.

[7] C. McCool et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *IEEE ICMEW*, 2012.

[8] J. L. Newman and S. J. Cox. Language Identification Using Visual Features. *IEEE Trans. on Audio, Speech, and Language Processing,*, 20(7):1936–1947, 2012.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on PAMI*, 2002.

[10] M. K. Omar and J. Pelecanos. A novel approach to detecting non-native speakers and their native language. In *IEEE ICASSP*, 2010.

[11] J. Orozco, O. Rudovic, J. Gonzàlez, and M. Pantic. Hierarchical On-line Appearance-Based Tracking for 3D Head Pose, Eyebrows, Lips, Eyelids and Irises. *Image and Vision Computing*, February 2013.

[12] G. Potamianos, H. P. Graf, and E. Cosatto. An image transform approach for HMM based automatic lipreading. In *IEEE ICIP*, 1998.

[13] E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak. Detecting nonnative speech using speaker recognition approaches. In *Proc. IEEE Odyssey Speaker and Language Recognition Workshop*, 2008.

[14] F. William, A. Sangwan, and J. H. L. Hansen. Automatic Accent Assessment Using Phonetic Mismatch and Human Perception. *IEEE Trans. on Audio, Speech & Lang. Proc.*, 2013.