Assessing the efficacy of benchmarks for automatic speech accent recognition

Benjamin Bock
Lawrence Technological University
21000 W Ten Mile Road
Southfield, MI 48075, USA
bbock@ltu.edu

Lior Shamir
Lawrence Technological University
21000 W Ten Mile Road
Southfield, MI 48075, USA
lshamir@mtu.edu

ABSTRACT

Speech accents can possess valuable information about the speaker that can be used in intelligent multimedia-based human-computer interfaces. The performance of algorithms for automatic classification of accents is often evaluated using audio datasets that include recording samples of different people, representing different accents. Here we describe a method that can detect bias in accent datasets, and apply the method to two accent identification datasets to reveal the existence of dataset bias, meaning that the datasets can be classified with accuracy higher than random even if the tested algorithm has no ability to analyze speech accent. We used the datasets by separating one second of silence from the beginning of each audio sample, such that the one-second sample did not contain voice, and therefore no information about the accent. An audio classification method was then applied to the datasets of silent audio samples, and provided classification accuracy significantly higher than random. These results indicate that the performance of accent classification algorithms measured using some accent classification benchmarks can be biased, and can be driven by differences in the background noise rather than the auditory features of the accents.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation* ; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Audio input/output*

General Terms

Reliability

1. INTRODUCTION

Accent classification is an important and increasingly relevant area of speech analysis [3] with application to multimedia data analysis, security biometrics, and auditory human-machine interface. The use of machine learning and signal processing algorithms has enabled automatic recognition of accents that can quantify and analyze traits of speech that may be inaudible or difficult for the human hearing to isolate. Effective automatic accent analysis can also further the field of speech and language cognition by allowing new possibilities such as quantitative accent relationships and speech mapping.

Research on automatic accent classification has been done in the past ∼25 years using numerous different approaches [2]. For instance, it has been found that automatic accent classification can be achieved by using source generator framework and analyzing prosodic features of speech samples [10]. Using an American English speech corpus with four accents, the classification rate was 81.5% when the speech text was unknown, and 88.9% with a known three-word set. That study also found that classification rate increases as the accent-sensitive word count increases [10]. Using the first order differences of Mel-cepstrum coefficients and energy, an isolated word and phoneme based-classification could achieve accuracy of 93% when using four different language accents and strings of seven to eight words [1].

Experiments with Australian English samples classifying between a native Australian accent, Lebanese Arabic accent, and a South Vietnamese accent showed 85.3% classification accuracy for accent pairs and 76.6.% accuracy for all three accents using accent-specific hidden Markov models (HMM's) and phoneme bigram language models [12]. By analyzing different positions within chosen syllables, accent classification of ∼93% was achieved between native Australian English and South Vietnamese Australian English, and 84% classification accuracy was achieved between native Australian English and Lebanese Arabic English speakers [4]. The same accents were also studied by [12], who reported on classification accuracy of 85.3% between pairs of accents, and other experiments suggest that machine analysis outperformed human listeners in the identification of British accents [9].

Further research on automatic accent classification focused on the ability to identify standard American English accents and Indian English accents using Gaussian mixture modeling, and showed that standard American English accent was correctly identified in accuracy of 85% while the accuracy of Indian English accent classification was 87.5% [7]. [17] reported on classification accuracy of 97.5% between Arabic

MOBIMEDIA 2015, May 25-27, Chengdu, People's Republic of China

DOI 10.4108/icst.mobimedia.2015.259033

English and Indian English accent pairs, and 95% when a tree-based learner and rule-based classifier was used [17]. It has been also found that for classifying a larger number of accents, heteroscedastic linear discriminant analysis (HLDA) and maximum mutual information (MMI) training can be used. In a dataset of 23 different foreign language accents in English, a detection rate of 32% percent was achieved [5].

However, as an emerging field of research it is important to validate the existing accent classification methods, and ensure that accent benchmark datasets are not biased. The performance of automatic accent recognition methods is often evaluated by the classification accuracy, which quantifies the ability of the method to associate audio samples with the correct accent.

Here we study audio datasets used for automatic accent classification, and show that classification accuracy far higher than random can be achieved by separating one second of silence from each audio sample. Our experiment suggests that experimental results evaluated using some speech corpora might be biased, but also provides another indication that benchmark datasets can be vulnerable to background noise, and should therefore be used with caution for assessing the actual performance of machine learning methods.

2. METHOD

Two audio analysis tools were used in this study. The first was jAudio, which is part of the jMIR open source music and audio analysis package [14]. jAudio extracts audio content descriptors that reflect various aspects of the sound such as 1D and 2D moments, area moments, spectral and harmonic spectral properties (flux, centroid, smoothness), beat histograms, zero crossing, Mel-Frequency Cepstral Coefficients (MFCC) and more, as described in [14]. A total of 78 numerical audio content descriptors were provided by jAudio. These features were classified by SVM (Support Vector Machine) using the SVM^{light} open source SVM implementation [11].

The second method was an audio analysis scheme used previously for automatic classification of whale sounds [23] and music [8]. The method first transforms each audio sample to its spectrogram, which is a 2D visualization of the audio such that the horizontal axis is the time, and the vertical axis is the frequency. The spectrograms are generated using the SoX (Sound Exchange), and then analyzed by the Wndchrm 2D analysis tool [22, 21], extracting a comprehensive set of 2883 2D numerical content descriptors from each spectrogram. The content descriptors include texture features such as Gabor, Haralick and Tamura textures, Radon transform features, Fractal features, Chebyshev Statistics, Multi-scale histograms, first four moments of the intensity values, edge features, statistics of the high-contrast 8-connected Otsu binary mask objects, Zernike features, and Chebyshev-Fourier features. A detailed description of the numerical descriptors is available in [18, 22, 16, 20, 21, 23].

After the 2D numerical content descriptors are computed, the feature vectors are classified using a WND (Weighted Nearest Distance) classifier [18], such that the Fisher discriminant scores of the features are used as weights, and 15% of the features with the highest Fisher discriminant scores are selected while the rest are rejected from the analysis [18, 22, 20, 21]. The method is described in detail in [18, 22, 16, 20, 21, 23]. All software used in this study is publicly available with open source.

Table 1: Number of samples per accent in the speech corpus

Accent	# speakers	# samples
American	438	1198
British	105	701
Canadian	93	679
Indian	54	422
Irish	8	80
Australian	35	305
New Zealand	14	140
South African	3	30
Total	751	3555

The method works by first separating a recording of silence from the beginning of the audio sample, before the person starts to speak. That ensures that no accent information exists in the audio samples, and the ability to classify them cannot be attributed to dataset bias. In this study the separation of the silent audio samples from the original samples was done using the SoX (SOund Exchange) open source software.

Two datasets were used in this study. The first speech corpus was VoxForge, obtained from www.VoxForge.org. VoxForge collects transcribed speech from volunteers and makes it available to open-source speech-recognition engines. The speech samples come from volunteers who visit the website, identify their accent, and record their voice while reading written prompts provided by the website. All volunteers are English speakers and are classified by accent (country of accent). The accents used in this experiment were American, Canadian, New Zealand, Australian, Indian, South African, and British.

The VoxForge dataset has been widely utilized in automatic speech analysis research, and also for the specific task of accent identification. For instance, it was used to show that an automatic accent analysis method could identify three accents pulled from the dataset with 80.1% accuracy [15], or to test a method of speaker identification using wavelet transform, entropy, standard deviation, and mean at the decomposition level [26]. Using 200 speakers, a 83.9% accuracy was achieved in identification of the speaker [26]. Another study used the dataset to utilize intra-modal fusion of multiple features from MFCC and wave transform, which produced higher results than single-feature methods [25]. A high classification rate of 98% between English and French samples was achieved using those accents from the VoxForge data, and 91% between the German and Italian accents [6].

Table 1 shows the number of samples and number of different speakers for each accent in the VoxForge speech corpus.

As mentioned above, we separated the first second from each sample, providing a dataset of samples such that each sample had one second of silence, and does not include any speech.

From each person we used one sample, so that our dataset had eight samples per accent.

The second dataset was The Speech Accent Archive [27], speech corpus comprised of English speakers from different countries and regions. All speakers read the same prompt – a simple message about picking up grocery items from the store. The entire corpus consisted of 152 different regions/accents. All speakers are recorded reading the following prompt: "Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

Since most of the samples did not contain one second of silence, just five samples from each of five different accents were used. The different accents are Portuguese, Dutch, English, French, and German.

3. RESULTS

Each class in the VoxForge dataset was separated to seven training samples and one test sample, and each experiment was repeated 20 times such that in each run different random samples were selected for training and test sets. Using the audio descriptors extracted from the spectrograms [23, 8], the average classification accuracy was 100%, showing that it was possible to classify the different accents when using just silent samples that do not contain any speech, and therefore do not contain accent information. When using the numerical audio descriptors extracted using jAudio the classification accuracy was also 100%. These results are in fact higher than previously reported accent classification methods that were tested with the full samples.

From the Speech Accent Archive we extracted one second of silence from the following accents: Portuguese, Dutch, English (American), French, and German. Since in many of the samples the beginning of the recording did not include a full second of silence, we were only able to extract one full second from five of the speakers in the five accents, and only four from Portuguese. Training with three samples and testing with two (one in the case of Portuguese), we received a classification accuracy of 28%. The classification accuracy of the Speech Accent Archive was much lower than the 100% accuracy observed with the VoxForge accent dataset, but it is still higher than the classification of random guessing, which is 20%.

The VoxForge dataset was also used for person identification by voice [13]. To test for possible bias in person identification by voice we attempted audio classification of the silent audio samples separated into classes such that each class contained the audio samples of a different speaker. The results showed that using 36 American speakers with 10 silent samples per person we observed classification accuracy of ∼82%. That classification accuracy is much higher than the ∼2.8% of random guessing. With 54 speakers and eight samples per person the classification accuracy was reduced to ∼79%, also far higher than the random guessing accuracy of ∼1.85%. With 75 speakers and five samples per person the classification accuracy was ∼72%.

4. CONCLUSION

The application of the method to two speech corpora show dataset bias, demonstrating that some speech corpora should be tested for possible bias to allow objective judgment of their ability to reflect the efficacy of speech recognition methods.

The results in this paper show that benchmarks for speech analysis should be collected such that the samples are normalized by the data acquisition session, hardware, etc'. For instance, if all samples of a certain class are acquired in the same session and then separated randomly to training and test sets, the samples can be matched by the acquisition session rather than by the content that the algorithms aim at analyzing. The same can happen if the samples for each class are collected at a different place, at a different time, by different hardware, or any other difference that might not be easily perceived manually, but can be sensed by computer algorithms and lead to overoptimistic classification results that do not reflected the actual performance of the algorithms.

5. DISCUSSION

Speech corpora are widely used for the development, testing, and evaluation of the performance of speech analysis systems. They have the advantage of providing objective comparison of the performance of different methods using the same data, and therefore allow a comparison in which the only variable is the pattern recognition method being tested. That experimental design makes it is possible to quantitatively compare the efficacy of speech analysis methods, and is widely used in machine learning and multimedia research also outside the scope of speech and language processing.

However, the high dimensionality of multimedia data makes it difficult to assess the ability of multimedia benchmarks to provide a reliable reflection of the problem at hand. Here we proposed a method that can identify the presence of dataset bias in speech recognition datasets, and showed that separating one second of silence from audio samples of accents provided high classification accuracy, even though the samples contained merely silence, and therefore contained no information about the accent. That showed that the classes could be identified by the background noise, which can be the result of the data acquisition process. For instance, in the case of VoxForge the samples of each speaker were collected from a different machine, using a different computer and audio hardware. Therefore, classification between individual speakers can be due to differences in the hardware used for acquiring the audio samples.

Background noise and artifacts in commonly used benchmarks were found to have a possible effect on the performance of pattern recognition methods tested using these datasets. Face images in face recognition benchmarks could be classified with high accuracy by using just background parts of the images that have no face or hair area in them, showing that the faces can be classified by artifacts, and not necessarily by the facial content [18]. Automatic analysis of microscopy images showed similar observation, where experiments with microscopy image datasets showed classification accuracy much higher than random even after all cell areas were removed from the images, and the images were classified without any cells in them [19]. A related observation showed that the classification of object recognition methods is more accurate when using training and test samples from the same benchmark, compared to using one benchmark for training and another for testing [24].

Since artifacts that differentiate between the classes are present in the background, it is reasonable to assume that such artifacts are also present in the foreground (cell regions of interest), so that segmentation or any other pre-processing cannot be safely used to correct for them.

6. REFERENCES

[1] L. M. Arslan and J. H. Hansen. Language accent classification in american english. *Speech Communication*, 18(4):353–367, 1996.

[2] W. Barry, C. Hoequist, and F. Nolan. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech & Language*, 3(4):355–366, 1989.

[3] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, et al. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786, 2007.

[4] K. Berkling, M. A. Zissman, J. Vonwiller, and C. Cleirigh. Improving accent identification through knowledge of english syllable structure. In *IEEE International Conference on Speech, Language and Signal Processing*, volume 98, pages 89–92, 1998.

[5] G. Choueiter, G. Zweig, and P. Nguyen. An empirical study of automatic accent classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4265–4268. IEEE, 2008.

[6] J. De Mori, M. Faizullah-Khan, C. Holt, and S. Pruisken. Spoken language classification. 2012.

[7] S. Deshpande, S. Chikkerur, and V. Govindaraju. Accent classification in speech. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pages 139–143. IEEE, 2005.

[8] J. George and L. Shamir. Computer analysis of similarities between albums in popular music. *Pattern Recognition Letters*, 45:78–84, 2014.

[9] A. Hanani, M. J. Russell, and M. J. Carey. Human and computer recognition of regional accents and ethnic groups from british english speech. *Computer Speech & Language*, 27(1):59–74, 2013.

[10] J. H. Hansen and L. M. Arslan. Foreign accent classification using source generator based prosodic features. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 836–839. IEEE, 1995.

[11] T. Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms.* Kluwer Academic Publishers, 2002.

[12] K. Kumpf and R. W. King. Automatic accent classification of foreign accented australian english speech. In *Fourth International Conference on Spoken Language*, volume 3, pages 1740–1743. IEEE, 1996.

[13] A. Maesa, F. Garzia, M. Scarpiniti, and R. Cusani. Text independent automatic speaker recognition system using mel-frequency cepstrum coefficient and gaussian mixture models. *Journal of Information Security*, 3(4), 2012.

[14] C. McKay. *Automatic music classification with jMIR.* PhD thesis, McGill University, 2010.

[15] G. Montavon. Deep learning for spoken language identification. In *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[16] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg. Wnd-charm: Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters*, 29(11):1684–1693, 2008.

[17] C. Pedersen. Accent classification from speech samples by use of machine learning. 2009.

[18] L. Shamir. Evaluation of face datasets as tools for assessing the performance of face recognition methods. *International Journal of Computer Vision*, 79(3):225–230, 2008.

[19] L. Shamir. Assessing the efficacy of low-level image content descriptors for computer-based fluorescence microscopy image analysis. *Journal of Microscopy*, 243(3):284–292, 2011.

[20] L. Shamir, S. M. Ling, W. W. Scott, A. Bos, N. Orlov, T. J. Macura, D. M. Eckley, L. Ferrucci, and I. G. Goldberg. Knee x-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2):407–415, 2009.

[21] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, 7(2):8, 2010.

[22] L. Shamir, N. Orlov, D. M. Eckley, T. Macura, J. Johnston, and I. G. Goldberg. Source code for biology and medicine. *Source Code for Biology and Medicine*, 3:13, 2008.

[23] L. Shamir, C. Yerby, R. Simpson, A. M. von Benda-Beckmann, P. Tyack, F. Samarra, P. Miller, and J. Wallin. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. *The Journal of the Acoustical Society of America*, 135(2):953–962, 2014.

[24] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011.

[25] G. K. Verma. Multi-feature fusion for closed set text independent speaker identification. In *Information Intelligence, Systems, Technology and Management*, pages 170–179. Springer, 2011.

[26] G. K. Verma and U. Tiwary. Text independent speaker identification using wavelet transform. In *International Conference on Computer and Communication Technology*, pages 130–134. IEEE, 2010.

[27] S. H. Weinberger and S. A. Kunath. The speech accent archive: towards a typology of english accents. *Language and Computers*, 73(1):265–281, 2011.