Computer and Human Recognition of Regional Accents of British English

Abualsoud Hanani, Martin Russell and Michael J. Carey

School of Electronic, Electrical and Computer Engineering, University of Birmingham, UK
{Aah648, M.J.Russell, M.Carey}@bham.ac.uk

10.21437/Interspeech.2011-281

Abstract

This paper is concerned with classification of the 14 regional accents of British English in the ABI (Accents of the British Isles) speech corpus. Results are reported using a state-of-the-art Language Identification system, variants of Huckvale's ACCDIST system, and human listeners. The best performance, 95.18% accuracy, is obtained using the text-dependent ACCDIST measure. The performance of a conventional (text-independent) acoustic Language Identification system is poor, but is improved significantly (89.6% accuracy) by the addition of phone sequence information. Human performance (58.25% accuracy) is much lower than expected.

Index Terms: Accent Recognition, ACCDIST, Accents of British Isles.

1. Introduction

Recent years have witnessed increased interest in automatic dialect and accent recognition from speech. This appears to be more difficult than language recognition, presumably because differences in morphology, lexicon, syntax, phonetics and phonology are typically more pronounced between languages than between dialects and accents of a language. There is some confusion in the speech technology literature between the terms 'dialect' and 'accent'. In British English 'accent' normally refers to systematic variations in pronunciation associated with particular geographic regions, while 'dialect' also includes the use of words that are characteristic words of those regions. So, for example, when a speaker from Yorkshire in the North of England pronounces "bath" to rhyme with "cat" rather than with "cart" he or she is exhibiting a Yorkshire (or at least northern English) accent, but use of the word "lug" to mean "ear" or "flag" to mean "paving stone" are examples of Yorkshire dialect [1]. Thus, the differences between accents are less than those between dialects. Speech technology is normally concerned with accent.

Accent recognition has many potential applications. Accent variation is a major source of variability for Automatic Speech Recognition (ASR), and recognizing the accent prior to ASR will enable the system to adapt this variation more effectively [2]. A better understanding of the acoustic-phonetic properties of accents might also lead to synthesis of regional accented speech [3]. This technology is also useful for recognizing a speaker's regional origin and his or her ethnicity [4].

Early work on accent recognition applied successful techniques from Language Identification (LID), such as Phone Recognition followed by Language Modelling (PRLM) [5] which exploited the phonotactic properties of the language or accent. Discriminative techniques provide further improvements to PRLM systems ([6], [2]). Modeling the acoustic properties of an accent with Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) has also achieved some success (e.g. [7]). However, some more recent research has exploited specific properties of accents. In [2] Biadsy et al. use the fact that, at least to a first approximation, accents share the same phone set, but the realization of these phones may differ. They reported improved performance compared with a conventional utterance level GMM-SVM system by using phone-dependent GMMs and creating 'supervectors' at the phone level. Huckvale in [3] took this a step further with his ACCDIST (Accent Characterization by Comparison of Distances in the Inter-segment Similarity Table) measure, by exploiting the fact that British English accents can be characterized by the similarities and differences between the realizations of vowels in specific words [8]. For example, for our speaker from Yorkshire the distance between the realizations of the vowels in "bath" and "cat" is small, but it is large between those in "bath" and "cart", whereas for a subject with a Southern English accent the opposite is the case. Huckvale reported an accent recognition accuracy of 92.3% on the 14 accents of British English in the ABI corpus [3].

2. Corpus description

The Accents of British Isles (ABI) speech corpus [9] was used in all experiments reported in this paper. The ABI speech recordings represent 13 different regional accents of the British Isles, plus standard British English. These were made on location in 13 different regions, namely Belfast, Birmingham, Burnley (Lancashire), Denbigh (North Wales), Elgin (Scottish Highlands), Dublin, Glasgow, Hull (East Yorkshire), Inner London, Liverpool, Lowestoft (East Anglia), Newcastle and Truro (Cornwall). In each case, twenty people were recorded (ten female and ten male) who were born in the region and had lived there for all of their lives. The standard southern British English speakers were selected by a phonetician. Each subject read twenty prompt texts.

The speakers were divided into three subsets; two with 93 and one with 94 speakers. Gender and accent were distributed equally in each subset. A "jack-knife" training procedure was used in which two subsets were used for training (train set) and the remaining subset for testing (test set). This procedure was repeated three times with different training and test sets, so that each ABI speaker was used for testing, and no speaker appeared simultaneously in the training and test sets. Overall performance was obtained from the combined scores over all test subsets.

Every participant read a short 92 word passage (the start of the "sailor" passage). The length of this recording varied from 30 to 45 seconds. The short passage recordings in the training set, with transcriptions aligned at the sentence level, were used to train the text-dependent ACCDIST systems (see section 3), and those in the test set were used for evaluation.

All recordings (including short passages) in the training set were used to train the text-independent systems (section 4). Two separate evaluations of the text-independent accent recognition systems were conducted, one using the 'short passage' utterances, and the other using 30-seconds cuts from all test recordings. In total, around 280 'short passage' utterances were used to evaluate the text-independent and text-dependent systems. The same utterances were also classified by human listeners. In addition, 1504 30-second cuts from all speakers in the test sets were used to evaluate the text-independent accent recognition systems, to enable comparison with standard language identification performance.

3. Text-dependent automatic systems

3.1. ACCDIST

Huckvale [3] exploits the fact that British English accents are characterized by similarities and differences between the realization of vowels in specific words (c.f. Wells' 'Key Words' [8]). His ACCDIST measure is less sensitive to speaker specific characteristics, other than accent, than measures that depend on absolute spectral properties.

The ABI corpus includes accent diagnostic material. A phonemic transcription of each of the short passage recordings was generated using standard British English pronunciations from BEEP dictionary [10], and force-aligned with the speech data using our British English phone recognizer (section 4.2). Our ACCDIST-based system differs from that in [3] in two ways. First, we used the 'short passage' rather than the 'short sentence' files in the ABI corpus. The 'short passage' was chosen because we believe it is more suitable for human perceptual experiments, and we wanted to use the same test material to test automatic and human recognition. Second, we used all of the 'short passage' recordings in our experiment, whereas only those recordings which were completed without errors or repetitions were used by Huckvale. This is because the method in [3] requires each file to have exactly the same phone-level transcription, so that the corresponding speaker distance tables are the same size. To obtain these tables, each realization of a vowel is split into two halves by time, and the average feature vectors (19 MFCCs plus energy) for each half are concatenated into a 40-dimensional vector. Then, distances are calculated between vectors from different contexts using an unweighted Euclidean distance [3].

In our system recordings need not correspond to exactly the same phone sequence. Instead, the speaker distance tables are built from vowel tri-phone segments ("phone-vowel-phone") rather than words. We also added vowel duration as an extra feature. For repeated tri-phones the mean feature vector was used. Hence, each 'short passage' recording is represented as a sequence of pairs (v_i , p_i), where v_i is the 41-dimensional feature vector of the i^{th} tri-phone and p_i is its label.

A speaker distance table was calculated for each speaker by finding the distances between the feature vectors of every tri-phone pair in the common tri-phones list. Then, the mean of the resulting speakers' distance tables was calculated for each accent. Accent recognition was used the correlation distance between the test speaker distance table and the accent mean distance tables, taking into account only those tri-phone-pairs which occurred in the test utterance. An obvious shortcoming of the vowel tri-phone approach is the limited vowel context, compared with the whole-word contexts in Huckvale's system.

3.2. ACCDIST with SVM

The success achieved by applying SVMs to supervectors constructed from stacked MAP-adapted GMM means in speaker and language recognition [7] motivated us to apply SVMs to the speaker distance tables in our ACCDIST-based system. In our version of Huckvale's system, the average of the speaker distance tables for a given accent was used to represent that accent. By contrast, in our ACCDIST-SVM system, SVMs were applied to the 'vectorized' speaker distance tables of all accents by labeling the distance tables of one accent as a target class (+1) and the remaining distance tables as a background class (-1). This results in one SVM for each accent. A test speaker distance table is evaluated against every accent model. The correlation distance kernel K was used in training and evaluating the SVM systems:

$$K(V_1, V_2) = \left[\frac{V_1 - \overline{V_1} I_J}{\sqrt{\sum_{j=1}^{J} (V_1^j - \overline{V_1})^2}} \right] * \left[\frac{V_2 - \overline{V_2} I_J}{\sqrt{\sum_{j=1}^{J} (V_2^j - \overline{V_2})^2}} \right]^T \qquad (1)$$

where, V_1, V_2 are the two distance table vectors. $\overline{V_1}$ and $\overline{V_2}$ are the mean of V_1 and V_2 respectively and I_J is a vector of 1s. $K(V_1, V_2)$ is the correlation distance kernel. For the 'short passage' 105 triphones were used, resulting in a 105 × 105 symmetric distance matrix. Hence, the dimension of the 'vectorized' speaker distance tables is $n(n-1)/2 = 5460$.

4. Text-independent automatic systems

4.1. Phonotactic systems

The success of the PRLM approach for language and dialect identification motivated us to apply it to our accent recognition task. In the PRLM approach, a sequence of phones is extracted from each training utterance of each accent using a single phone recognizer. An n-gram language model is trained on the resulting phone sequences using Support Vector Machines (SVM), one SVM for each accent. Before building the language models, a weighting technique proposed in [11] and used in our language ID system in [12] is applied to the n-gram probabilities in order to emphasize the most discriminative components (i.e. those which are common in one accent but not in others). This weighting also de-emphasizes the n-gram components that are common in all accents, as they do not carry useful discriminative information. The weight w_j for component C_j is given by:

$$w_j = g_j \left(\frac{1}{p(C_j / All)} \right) \qquad (2)$$

Where g_j is a function used to smooth and compress the dynamic range (for example, $g_j(x) = \sqrt{x}$) and $p(C_j / All)$ is the probability of n-gram component C_j across all accents. The components which have zero occupancy in all accents are removed since they do not carry any useful information. A benefit of discarding these low-occupancy components is that it reduces the feature dimension dramatically, particularly for the high order n-gram systems.

In recognition, a phone sequence is extracted from the test utterance; an n-gram probability vector is computed and weighted with the weight factor above. Then the weighted n-gram vector is evaluated using the SVMs for the 14 accents.

Using Parallel PRLM (PPRLM) with multiple phone recognizers trained on different languages and combining them in the back end improves the performance of language, dialect and accent ID systems, e.g. [5].

4.2. Phone recognizers

In our phonotactic systems, we used three different phone recognizers for Czech, Hungarian and Russian from a toolkit developed by Brno University of Technology1. These were

[1] www.fit.vubr.cz/research/groups/speech/sw/phnrec

trained on the SpeechDat-E databases using a hybrid approach based on Neural Networks and Viterbi decoding [13].

Since British accent identification is our goal, we hypothesized that a British English phone recognizer would be advantageous. Therefore, we built a British English decision-tree triphone-based phone recognizer, using the HMM toolkit (HTK)[2]. We extracted 12-dimensional Perceptual Linear Predictive (PLP) features from 25 ms frames, with a frame shift of 10 ms. Each feature vector is 39 dimensional, comprising 13 features (12 PLP features plus energy), plus 13 "delta", and 13 "double-delta" parameters. The features are normalized using cepstral mean normalization. We used training data in the ABI corpus (section 2) to train acoustic models. All phone HMMs consist of 3 states without state-skipping, with one 16 component GMM per state. We trained a bigram phone-level language model on the ABI training set. The pronunciation dictionary was generated from the British English pronunciation dictionary (BEEP) [10].

4.3. Acoustic systems

Modeling acoustic features is an alternative method that has been successfully used in language, dialect and accent recognition systems. Most acoustic-based approaches use GMMs to model distributions of acoustic feature vectors. The most successful variants are GMM-UBM, GMM-SVM and GMM-*n*-gram [14], and these are described below.

In the GMM-UBM approach, a Universal Background Model (UBM) was built using utterances from the training sets of all accents. Accent-dependent models are obtained by MAP adaptation (adapting means and weights) of the UBM, using the accent-specific enrollment data. The result is one UBM and fourteen accent-dependents GMMs.

In our GMM-SVM system, the speech data from each individual speaker was used to estimate the parameters of a GMM by MAP adaptation of the UBM. The adapted GMM mean vectors are then concatenated into a 'supervector', and the accent classes are assumed to be linearly separable in this supervector space. The supervectors are used to build one SVM for each accent, by treating that accent as the 'target' class and the others as the 'background' class.

In the third acoustic based system, GMM-*n*-gram, the UBM GMM which was trained on the training data for all accents was used as a tokenizer to generate sequences of GMM component indices from the sequence of cepstral features. The resulting sequence is used to train an *n*-gram language model for each accent using SVMs. Compared with the PRLM system described earlier, the phone recognizers are replaced by an accent-independent GMM which produces a sequence of Gaussian component indices instead of a sequence of phones. The other parts of these two types of system are the same, including the use of discriminative weighting to emphasize the GMM components which represent the accent specific features and de-emphasize the components which represent features that are common in all accents.

In all of our systems, the score S_j for each accent model is normalized using the highest competing score for the other accents (max-log-likelihood score normalization):

$$s'_j = s_j - \max_{i \neq j}(s_i) \qquad (3)$$

4.4. Phonotactic-acoustic fusion

Using 30-sec testing cuts, the outputs of the four acoustic systems (GMM-UBM, GMM-SVM, GMM-uni-gram and GMM-bi-gram) were fused together using Brummer's multi-class linear logistic regression (LLR) toolkit[3] (row 3, Table 1). The four acoustic systems were also fused with 12 phonotactic systems; (2, 3 and 4)-gram with four different phone recognizers (row 4, Table1). The same fusion process was repeated for the "Short Passage" testing files (row 5, Table1).

Because there is no development set to train the logistic regression fusing coefficients, we divided the testing speakers in each 'jack-knife' round into two sets. The accent and gender of speakers are distributed equally in both sets. One set is used to find the coefficients for fusing the systems on the second set, and *vice versa*.

5. Human experiment

To provide a baseline against which the automatic accent recognition systems could be compared, a web-based human perceptual experiment was conducted using exactly the same 'ShortPassage' test recordings. Twenty four native British English speaking subjects aged between 21 and 78 each listened to sets of 20 recordings selected randomly from the test set. For each recording, subjects were asked to identify the accent of the speaker (out of 14 accents), the speaker's gender and age and to state their confidence in their decisions. Human listeners scored an average error rate of 41.75% for the accent classification task.

6. Experimental setup

All speech signals were bandpass filtered (0.23 - 3.4 KHz) to simulate a telephone channel, and downsampled to 8 KHz.

As indicated previously, three different orders of *n*-grams (*n*=2, 3 and 4), are used to model the phone sequences from the four phone recognizers (Section 3.1). These result in twelve PRLM systems: (4 phone recognizers) × (3 *n*-gram systems).

In the acoustic systems, the feature vectors are based on nineteen cepstral coefficients derived from the power output of nineteen Quadrature pairs of linear phase FIR filters. Periods of silence were discarded using a pitch-based voice activity detector. The Mel Frequency Cepstral Coefficients (MFCC), including C_0, are concatenated with Shifted-Delta Cepstra coefficients (SDC) with a 7-3-1-7 configuration [15], giving a total of 68 features per frame at a frame rate of 100 frame per second. RASTA filtering is applied to the power spectra, and feature warping, with 3-seconds windows, is applied to the final feature vectors.

One UBM GMM, with 4096 mixture components, is trained on the acoustic training data with 5 iterations of E-M, updating all parameters; means, diagonal covariances and weights. Fourteen accent-dependent GMMs are MAP-adapted from the UBMs using accent specific data. The UBM GMM means are also MAP adapted using speech data from each speaker for each accent, generating the GMM supervectors which are used to train the GMM-SVM system (section 4.3).

The same UBM GMM (4096 components) was also used to produce a sequence of GMM component indices (the best component in the case of the bi-gram model, and all components with probabilities for the uni-gram model). The uni-gram and bi-gram systems model the output on the GMM-tokenizer with an SVM.

All of the SVM models in this paper were trained and evaluated using the free SVM-KM SVM MATLB toolbox[4].

[2] htk.eng.cam.ac.uk

[3] Niko.brummer.googlepages.com/focalmulticlass

[4] Asi.insa-rouen.fr/enseignants/~arakotom/toolbox

The acoustic experiments were run on an Nvidia Geforce GTX260 Graphics Processing Unit (GPU), comprising 216 floating-point processors and 1.76GB RAM, and an Nvidia C106 Tesla machine with approximately similar performance. Programming was done in MATLAB, GPUmat and CUDA.

7. Results and discussion

The text-independent accent recognition systems applied in this paper are considered to be state of the art. The acoustic based systems were previously evaluated on the NIST 2003 Language Recognition Evaluation (LRE) test, and the average equal error rate (EER), estimated using 1200 30-sec testing utterances from 12 different languages, is 0.83% (row 2, Table 1) [12]. For Accent Identification (AID), the same acoustic systems score 12.33% EER on the 14 ABI accents (row 3, Table 1), which is low compared with LID, especially since the ABI data was not recorded over a telephone channel. It seems that the differences between the inventories of sounds in different accents in the same language are less pronounced than those between different languages. Another possible reason is the different number of classes and the different amounts of training data in the Callfriend[5] and ABI corpora. However, the average EER of the text-independent system is improved by 48% when it is fused with the phonotactic systems (row 4, Table 1). This improvement implies that the phonotactic and acoustic systems are complimentary, capturing different features for accent classification.

Table 1: *Average EER [%] and Accuracy [%] of the described systems*

System	EER [%]	Acc. [%]	90% conf.	No. tests
Acoustic LID on NIST 2003 30-sec cuts	0.83	97.3	±0.27	1200
Acoustic AID on 30-sec cuts	12.33	73.6	±0.92	1504
Acoustic-Phonotactic AID on 30-sec cuts	6.4	88.8	±0.69	1504
Acoustic-Phonotactic AID on “ShortPassage” recordings	4.52	89.6	±1.43	280
ACCDIST-Correlation Distance AID on “ShortPassage” recordings	2.66	93.17	±0.98	280
ACCDIST-SVM AID on “ShortPassage” recordings	1.87	95.18	±0.83	280

To be able to compare the text-dependent and text-independent accent recognizers directly, we evaluated them both on the same utterances (the 30 – 45 second ‘ShortPassage’ recordings, see section 2). The EER for the text-independent systems (row 5, Table 1) is reduced by a further 29.3% because of the longer utterances. Both of the text-dependent ACCDIST-based systems outperform the text-independent systems, confirming the utility of utilizing the differences between vowel realizations in specific contexts. The EER of the ACCDIST system proposed by Huckvale [3] is improved by 29.7% with an SVM with a correlation distance kernel (row 7, Table 1).

The human accent recognition experiment, which used exactly the same test material as the automatic systems, resulted in an accent recognition accuracy of 58.25%. This is much lower that any of the automatic systems.

8. Conclusions

In this paper we have investigated the effectiveness of standard techniques from Language Identification for Accent Identification on 14 regional accents of British English. We have also compared the performance of these systems with variations of the text-dependent ACCDIST-based system, proposed by Huckvale. Using knowledge of the transcription of speech, the ACCDIST-based approach achieved a 41.1% reduction in EER and a 4% increase in accuracy relative to the best of the LID system. A further EER improvement of 29.7% and 2.15% increase in accuracy was obtained by classifying the ACCDIST tables using an SVM with correlation distance kernels. The automatic systems (LID and ACCDIST) have been found to perform much better than the human listeners.

9. References

1. Elmes, S., “Talking for Britain: A Journey through the Nation’s Dialects”, Penguin Books, 2005.
2. Biadsy, F., Hirschberg, J., Collins, M., "Dialect Recognition using Phone-GMM-Supervector-based SVM Kernel", in Pro. InterSpeech. 2010.
3. Huckvale, M., "ACCDIST: An Accent Similarity Metric for Accent Recognition and Diagnosis", in Speaker Classification II, Lecture Notes in Computer Science, 2007. Volume 4441.
4. Hanani, A., Russell, M., and Carey, M., “Speech-Based Identification of Social Groups in a Single Accent of British English by Humans and Computers”, to appear in Proc. IEEE ICASSP 2011: Prague.
5. Zissman, M. A., "Comparison of four approaches to automatic language identification of telephone speech". IEEE Trans. of Sp and Audio Proc., 1996. **4**(1): p. 31-44.
6. Richardson, F. S., Campbell, W.M., and Torres-Carrasquillo, P.A., "Discriminative N-Gram Selection for Dialect Recognition", Interspeech 2009, Brighton, UK., Sept. 6, 2009.
7. Campbell, W.M., Sturim, D.E., and Reynolds, D.A., "Support vector machines using GMM supervectors for speaker verification". Signal Processing Letters, IEEE, 2006. **13**(5): p. 308-311.
8. Wells, J.C., “Accents of English, volume 2 The British Isles”: Cambridge University Press, 1982
9. D'Arcy, S.M., Russell, M.J., Browning, S.R. and Tomlinson, M.J, "The Accents of the British Isles (ABI) Corpus". Proc. Modelisations pour l'Identification des Langues, MIDL Paris, 2005: p. 115-119.
10. "ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz", [cited March 30, 2011].
11. Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A., Leek, T. R.,“Phonetic speaker recognition with support vector machines”, in Advances in Neural Information Processing Systems 16, 2004.
12. Hanani, A., Carey, M., and Russell, M., “Improved Language Recognition using Mixture Component Statistics”, Proc. Interspeech 2010, Tokyo, Japan, September 2010.
13. MATĚJKA, P., SCHWARZ, P., ČERNOCKÝ, J., CHYTIL, P., "Phonotactic Language Identification using High Quality Phoneme Recognition", in Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology. 2005: Lisbon. p. 2237-2240.
14. Torres-Carrasquillo, P., Singer, E., Campbell, W. M., Gleason, T., McCree, A., Reynolds, D. A., Richardson, F., Shen, W., Sturim, D., "The MITLL NIST LRE 2007 Language Recognition System", in Interspeech'08. 2008: Brisbane, Australia.
15. Torres-Carrasquillo, P. A., E.S., M.A. Kohler, R.J. Greene, D.A Reynolds, and J.R. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features". Proc. ICSLP 02, 2002: p. 89-92.

[5] www.ldc.upenn.edu/catalog