Computer and Human Recognition of Regional Accents of British English.

Abualsoud Hanani
Birzeit University
9 PUBLICATIONS **61** CITATIONS
SEE PROFILE

Martin J Russell
University of Birmingham
125 PUBLICATIONS **1,360** CITATIONS
SEE PROFILE

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 59–74

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Human and computer recognition of regional accents and ethnic groups from British English speech[☆]

A. Hanani, M.J. Russell*, M.J. Carey

School of Electronic, Electrical and Computer Engineering, University of Birmingham, Birmingham B15 2TT, UK

Received 29 April 2011; received in revised form 18 December 2011; accepted 8 January 2012
Available online 17 January 2012

Abstract

The paralinguistic information in a speech signal includes clues to the geographical and social background of the speaker. This paper is concerned with automatic extraction of this information from a short segment of speech. A state-of-the-art language identification (LID) system is applied to the problems of regional accent recognition for British English, and ethnic group recognition within a particular accent. We compare the results with human performance and, for accent recognition, the 'text dependent' ACCDIST accent recognition measure. For the 14 regional accents of British English in the ABI-1 corpus (good quality read speech), our LID system achieves a recognition accuracy of 89.6%, compared with 95.18% for our best ACCDIST-based system and 58.24% for human listeners. The "Voices across Birmingham" corpus contains significant amounts of telephone conversational speech for the two largest ethnic groups in the city of Birmingham (UK), namely the 'Asian' and 'White' communities. Our LID system distinguishes between these two groups with an accuracy of 96.51% compared with 90.24% for human listeners. Although direct comparison is difficult, it seems that our LID system performs much better on the standard 12 class NIST 2003 Language Recognition Evaluation task or the two class ethnic group recognition task than on the 14 class regional accent recognition task. We conclude that automatic accent recognition is a challenging task for speech technology, and speculate that the use of natural conversational speech may be advantageous for these types of paralinguistic task.

Keywords: Accent identification; British English; Language identification; Gaussian Mixture Model; Support Vector Machine

1. Introduction

A speech signal contains a wealth of information over and above its linguistic content, including clues to the geographical, social and ethnic background of the speaker. In the case of British English, most native listeners would be more or less aware of the speaker's regional accent, and a listener from the same region might also be aware of the speaker's social or geographical 'subgroup' within the region. In the first volume of "Accents of English", Wells defines 'accent of English' as "a pattern of pronunciation used by a speaker for whom English is the native language or, more generally, by the community or social grouping to which he or she belongs" (Wells, 1982). This is different from 'dialect' which also includes the use of words that are characteristic of those regions. So, for example, when a

[☆] This paper has been recommended for acceptance by 'Dr. Anton Batliner'.

* Corresponding author. Tel.: +44 121 414 3093; fax: +44 121 414 4291.
E-mail addresses: Aah648@bham.ac.uk (A. Hanani), m.j.russell@bham.ac.uk (M.J. Russell), m.carey@bham.ac.uk (M.J. Carey).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.01.003

speaker from Yorkshire in the North of England pronounces *bath* with the same vowel quality as *cat* rather than *cart* he or she is exhibiting a Yorkshire (or at least north of England) accent, but use of the word *lug* to mean “ear” or *flag* to mean “paving stone” are examples of Yorkshire dialect (Hughes et al., 2005; Elmes, 2005).

Automatic accent recognition from speech has a number of potential applications. Accent is a major source of variability for automatic speech recognition (ASR) (Humphries and Woodland, 1997; Tjalve and Huckvale, 2005; Biadsy et al., 2010), and recognizing a speaker’s accent prior to ASR could enable a system to accommodate this variation more effectively, for example by choosing appropriate acoustic and lexical models. However, even within a homogeneous population of subjects who were born in the same town or city and have lived there all of their lives, and therefore notionally speak with the same regional accent, there are likely to be significant variations. Typically these will include speakers whose accent is close to standard British English and other variations in accented speech associated with different social, geographical or ethnic groups. Therefore a more interesting challenge is to develop a continuous space representation of speakers and accent, such that subjects who are close in this space speak in a similar manner and, from the perspective of automatic speech recognition, can be characterized by similar sets of model parameters.

Coupled with a capability to synthesize regionally accented speech (for example, Yamagishi et al., 2010), automatic accent recognition could also be used to select appropriately accented synthetic speech in the context of an interactive dialogue system.

Much of the existing work on automatic accent recognition takes its lead from language identification (LID), and as in LID the different approaches can usefully be partitioned into acoustic methods, which exploit differences between the distributions of sounds in different accents, and ‘phonotactic’ approaches which exploit accent-dependent differences in the sequences in which these sounds occur (Zissman, 1996). An early example of the latter is Zissman’s work (Zissman et al., 1996) on the application of phone recognition followed by Language Modelling (PRLM) to accent recognition. The performance of PRLM can be further improved by the use of discriminative methods that focus on phones or phone sequences that are characteristic of an accent (Richardson et al., 2009; Biadsy et al., 2009). Another technique borrowed from LID is the use of Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) to model the acoustic properties of accented speech, and this has also achieved some success (for example, Richardson et al., 2009). Other acoustic-based approaches include the use of phone durations and average cepstra (Miller and Trischitta, 1996), phone and word-level Hidden Markov Models (HMMs) (Arslan and Hansen, 1996; Teixeira et al., 1997; Lincoln et al., 1998; Huang et al., 2007), and stochastic trajectory models (Angkititrakul and Hansen, 2006).

Some more recent research has exploited specific properties of accents. The approach described in Biadsy et al. (2010) uses the fact that, at least to a first approximation, accents share the same phoneme inventory, but the realization of these phonemes may differ. They report improved performance compared with a conventional utterance level GMM–SVM system using phoneme-dependent GMMs and creating ‘supervectors’ at the phoneme level. Huckvale (2007) 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 (Wells, 1982; Barry et al., 1989). 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-1 corpus (Huckvale, 2007).

Studies of human accent recognition, and the effects of the linguistic backgrounds of the listeners involved, have been reported for six regional accents of French (Woehrling and Boula de Mareüil, 2006), three regional accents of British English (Ikeno and Hansen, 2006), and non-native accented English (Arslan and Hansen, 1996). The ability of human listeners to distinguish between African American English and Standard American English has also been studied (Purnell et al., 1999; Walton and Orlikoff, 1994).

The comparison of different approaches to accent recognition is difficult because of the absence of standard corpora or evaluation methodologies. Corpora that have been used include various collections of non-native English (Arslan and Hansen, 1996; Teixeira et al., 1997; Angkititrakul and Hansen, 2006; Huang et al., 2007), British and American English (Lincoln et al., 1998), different corpora of regional accents of British English (D’Arcy et al., 2005; Ikeno and Hansen, 2006; Huang et al., 2007; Huckvale, 2007), five varieties of spoken Arabic (Biadsy et al., 2010), and six regional French accents (Woehrling and Boula de Mareüil, 2006).

In this paper we apply a state-of-the-art LID system to extract two types of paralinguistic information from English[1] speech, namely the speaker's regional accent and, in the case of Birmingham accented speech, the ethnic group to which the speaker belongs. As well as measuring the overall performances of our LID system, we report the performance of its acoustic and phonotactic subsystems on these tasks. These are compared with human listener performance, and in the case of regional accent recognition, with two systems based on Huckvale's ACCDIST measure (Huckvale, 2007).

An objection to this approach is that the application of LID techniques to these problems is counter-intuitive, because regional accent and ethnicity within a given language are examples of variability to which an LID system should, by definition, be insensitive. However, LID systems are based on generic statistical and pattern recognition methods, primarily GMMs, SVMs and *n*-gram language models. These are supplemented by normalization techniques designed to remove irrelevant variations from the sequences of feature vectors that are to be classified, where the definition of 'irrelevant' depends on the classification task. For example, in LID, a technique such as inter-session variability compensation (Vair et al., 2006) seeks to accommodate all differences between utterances from the same language (including speaker, regional accent and ethnic group differences), whereas for accent recognition the same technique is only trained to accommodate differences between utterances representing the same regional accent. Our premise is that within a given language (in this case English) the distributions of acoustic feature vectors, or phone *n*-grams, corresponding to different regional accents are sufficiently distinct to enable these methods to be applied successfully to regional accent recognition. The systematic differences between various regional accents of British English are well documented (Wells, 1982; Hughes et al., 2005). For example, according to Wells (1982) the "two most important characteristics setting northern accents apart from southern ones are (i) the absence of the *foot–strut* split, i.e. the lack of a phonemic opposition between the vowels of *foot* and *strut*; and (ii) the absence of *bath* broadening, i.e. the use in *bath* words of the vowel of *trap*" (here 'northern accent' refers to an accent associated with the north or midlands of England). Of course, both northern and southern accents include both *bath* vowels. However, their relative frequencies and precise acoustic realizations in different accents are unlikely to be the same, and this, in principle, is sufficient information for recognition. A quantitative analysis of the vowel systems for each of the 14 accents in the ABI-1 corpus is presented in D'Arcy (2007), which shows scatter plots of the first formant frequency F_1 against the difference between this and the second formant frequency, $F_2 - F_1$, for 11 monophthong vowels for each accent. These diagrams show both systematic inter-accent differences, and considerable intra-accent consistency. Accent differences in British English are not restricted to vowels (Wells, 1982), however quantitative data is less readily available.

In fact, there are a number of precedents in speech pattern classification where methods that give a performance gain for one task are also advantageous for another, complimentary task. For example, the delta-cepstrum was first proposed as an additional set of acoustic features for *speaker* recognition (Furui, 1981) but now also features in most automatic *speech* recognition systems. However, this is not our main argument.

The objective of the work described in this paper is to determine whether or not these inter-accent differences, and the differences between the speech of subjects from different ethnic groups, are sufficient to enable methods from LID to be applied successfully to automatic accent and ethnic group recognition.

The paper is organized as follows. In Section 2 we describe the corpora that are used in our experiments. The first corpus, the NIST 2003 Language Recognition Evaluation (Section 2.1) is a standard benchmark that we use to demonstrate the efficacy of our LID system. Our system achieves an Equal Error Rate (EER) of 0.7% on this task, which is comparable with the best published result. For regional accent recognition we use the 14 accents in the ABI-1 "Accents of the British Isles" corpus (Section 2.2). Because this is a transcribed corpus, we are able to compare our text-independent LID system with text-dependent approaches based on ACCDIST, as well as with human listeners. For our final study we use the "Voices across Birmingham" (VaB) corpus of telephone conversational speech (Section 2.3). The two largest ethnic groups in Birmingham (UK) are the 'Asian' and 'White' communities, and the VaB corpus includes a substantial amount of data from each group. We attempt to assign a subject to one of these two groups using a 40 s sample of his or her speech. Although we refer to this as 'ethnic group' classification, and to the two groups as 'Asian' and 'White', it is clear that we are actually concerned with differences between the patterns of pronunciation and language usage between the two communities, and not explicitly with ethnicity. We compare the performance of our LID system with that of human listeners on this task. Related human perceptual studies for varieties of American English are reported in Walton and Orlikoff (1994) and Purnell et al. (1999).

[1] Unless otherwise stated, in this paper 'English' refers to 'British English' speech.

Table 1
Summary of speech corpora used in the studies (Conv. = conversational speech).

Corpus	CallFriend	NIST LRE 2003	ABI-1	VaB
Style	Conv.	Conv.	Read	Conv.
Channel	Telephone	Telephone	Head mic	Telephone
Sample rate	8 kHz	8 kHz	22.05 kHz (resamp. 8 kHz)	8 kHz
Classes	12 Languages	12 Languages	14 Accents British English	2 Ethnic groups
Use	LID training	LID testing	Accent ID training and testing	Ethnic group ID training and testing

Section 3 describes the acoustic and phonotactic components of our LID system and our ACCDIST-based systems, Section 4 describes our experiments with human listeners, and Section 5 gives more information about the experimental procedure. The results are presented and discussed in Section 6. In summary, for the 14 regional accents of English in the ABI-1 corpus (good quality read speech), our LID system achieves a recognition accuracy of 89.6%, compared with 95.18% for our best ACCDIST-based system and 58.24% for human listeners. Our LID system distinguishes between the two Birmingham ethnic groups with an accuracy of 96.51% compared with 90.24% for human listeners. Our conclusions are presented in Section 7.

2. Speech corpora

The speech corpora used in this work are summarized in Table 1.

2.1. CallFriend and NIST2003

Our LID system is validated on the standard NIST 2003 Language Recognition Evaluation (LRE) closed-set task. The training data consists of the train set and the development set of the CallFriend corpus.[2] Each set contains 20 half-hour two-sided telephone conversations for each of 15 languages and dialects. The 12 languages are; English, Arabic, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. English, Spanish and Mandarin have two dialects. The NIST 1996 evaluation set (lid96e1) is used to train the back-end for calibration and fusion. The NIST 2003 evaluation set (lid03e1) contains test utterances of 30 s, 10 s and 3 s segments, but only 30 s segments are used in our evaluation. This evaluation subset consists of 1280 utterances; 80 for each language come from the CallFriend corpus, and 160 for English and 80 for Japanese come from the Callhome corpus.

2.2. The “Accents of the British Isles” (ABI-1) corpus

The Accents of British Isles (ABI) speech corpus (D’Arcy et al., 2005) was used in all of our regional accent recognition experiments. The ABI-1 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, 20 people were recorded (normally 10 women and 10 men) who were born in the region and had lived there for all of their lives. The standard southern English speakers were selected by a phonetician. Each subject read 20 prompt texts, ranging from ‘task oriented’ texts which are representative of generic applications of automatic speech recognition, to ‘phonetic’ texts chosen for their phonetic content. The later includes the “Sailor Passage” (SCRIBE, 1998), which was split into three parts of approximately equal length. The first section, referred to as SPA (Sailor Passage A) comprised 92 words and its recordings varied between 30 and 45 s in duration. Word-level transcriptions, aligned at the sentence level, are available for all of the ABI-1 recordings.

The ABI-1 recordings were made using head mounted and desk microphones, and sampled at 22.05 kHz. For the accent recognition experiments reported here, the head-mounted microphone recordings were bandpass filtered (0.23–3.4 kHz) to simulate a telephone channel, and downsampled to 8 kHz. The speakers were divided into three

[2] CallFriend Corpus: http://www.ldc.upenn.edu/Catalog.

subsets; two with 93 and one with 94 speakers. Gender and accent were distributed equally in each subset. A "jackknife" training procedure was used in which two subsets were used for training and the remaining subset for testing. This procedure was repeated three times with different training and test sets, so that each ABI-1 speaker was used for testing, and no speaker appeared simultaneously in the training and test sets.

Two separate evaluations of the text-independent accent recognition systems were conducted, one using 30-s extracts from all test recordings, and the other using the SPA utterances. The first test set, 1504 30-s extracts from all speakers in the ABI-1 test sets, was used to enable comparison with standard language identification performance on the NIST 2003 evaluation set (where we also used 30 s test utterances). The second test set, comprising approximately 280 SPA utterances, was used to evaluate and compare the text-independent and text-dependent automatic systems, and human listeners on the accent identification task.

2.3. The "Voices across Birmingham" (VaB) corpus

The "Voices across Birmingham" (VaB) corpus was used in the "ethnic group" recognition experiments. The goal of the VaB project is to capture variations in conversational speech across the people of the city of Birmingham in the UK. It currently comprises approximately 175 h of recordings of telephone conversational speech between participants who were born in or around the city. Each participant made up to 1 h of free telephone calls, which were routed through an Aculab Prosody X telephony card for automatic recording. Both participants in the call were aware that they were being recorded and of the purpose of the recording.

Significant immigration into Birmingham from Asia began in the 1960s. According to the 2001 census of England and Wales, which included questions about the ethnicity of residents,[3] approximately 70.4% of Birmingham's population categorized themselves as 'White' and 19.5% as 'Asian'. Twenty-nine percent of Birmingham's British Asian population gave their ethnicity as Indian, 53% as Pakistani, 11% as Bangladeshi and the remaining 7% as 'Other Asian'. The VaB project asked its participants similar questions about ethnicity. For the 'White' and 'Asian' groups there is sufficient data to conduct an experiment to study whether or not an individual can be classified automatically into the correct ethnic group from his or her speech (the VaB corpus does not distinguish between the different ethnic subgroups groups within the Asian community).

The Asian group can be further sub-divided into those subjects who were born in Birmingham (second generation) and those who were not. Only recordings from White and second generation Asian participants were included in the current experiments. The recordings from these two groups were divided into training and test sets. The training set consists of recordings from 242 different speakers (165 Asian and 77 White). The test set consists of 315 utterances from different speakers, each with maximum duration of 40 s. Of these, 175 are Asian (69 male, 106 female) and 140 are White (53 male, 87 female).

3. Automatic classification systems

It is convenient to divide the automatic systems used in this study into text-dependent systems, which require a word-level text transcription of the test utterance, and text-independent systems, which do not. The text-dependent systems are variants of Huckvale's ACCDIST approach (Huckvale, 2007). The text-independent systems are the components of our LID system, and can be further divided into phonotactic systems, which exploit phone sequence information for classification, and acoustic systems, which characterize the distribution of acoustic feature vectors for a particular class of speech. All of the systems are described below.

3.1. Front-end signal processing

The first stage in any speech pattern classification process is to convert the speech waveform into a sequence of acoustic feature vectors.

In our LID (Section 3.2.2) and ACCDIST-based systems (Section 3.3.1) the feature vectors are based on 19 Mel Frequency Cepstral Coefficients (MFCCs), including C0, derived from the log power output of 19 In-phase and

[3] 2001 Population Census in Birmingham (http://www.birmingham.gov.uk/community).

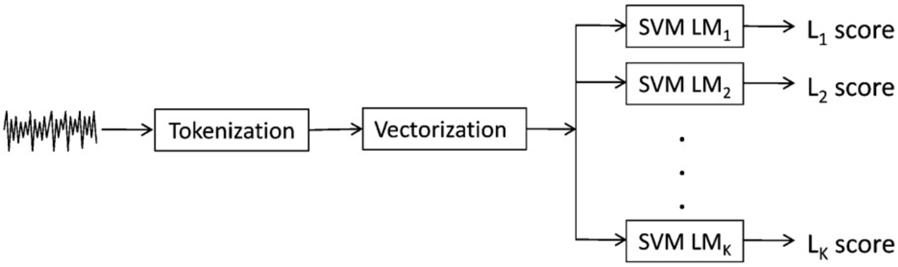

Fig. 1. A simple PRLM LID system (K is equal to the number of classes).

Quadrature pairs of linear phase FIR filters. These 19 coefficients plus energy form the (20 dimensional) feature vectors for the ACCDIST-based systems. In the other, text-independent acoustic systems, periods of silence are discarded using a pitch-based voice activity detector. The MFCCs, including C0, are concatenated with Shifted-Delta Cepstra coefficients (SDC) with a 7-3-1-7 configuration (Torres-Carrasquillo et al., 2002), giving a total of 68 features per frame at a frame rate of 100 frame per second. RASTA filtering (Hermansky and Morgan, 1994) is applied to the power spectra, and feature warping (Pelecanos and Sridharan, 2001), with 3s windows, is applied to the final feature vectors.

In our phonotactic systems (Section 3.2.1) the type of feature vector is dictated by the requirements of the phone recognizer. For our ‘in-house’ British English system 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.

3.2. *Text-independent automatic systems*

3.2.1. *Phonotactic systems*

In what follows, “class” refers to language, accent or ethnic group, as appropriate.

The success of the PRLM approach for language and dialect identification motivated us to apply it to our accent and ethnic group recognition tasks. A general PRLM system is depicted in Fig. 1. First, the ‘tokenizer’ converts the speech waveform into a sequence of symbols. In a conventional phonotactic LID system the symbols are phones and the tokenizer is a phone recognizer, but other configurations are possible (for example, the GMM-*n*-gram systems in Section 3.2.2). In ‘vectorization’ this sequence is used to estimate a vector of probabilities of a predefined set of symbol *n*-grams for each utterance. Each probability may be weighted according to the utility of the corresponding *n*-gram for classification. Finally, a set of SVM *n*-gram language models is trained on the vectors from utterances in the training sets, using one SVM for each class (Zhai et al., 2006).

In our phonotactic LID system (Hanani et al., 2010) the weighting technique proposed in Campbell et al. (2004) is applied to the *n*-gram probabilities in order to emphasize the most discriminative components (i.e. those which are common in one class but not in others), and de-emphasise the *n*-gram components that are common in all classes. The weight w_j for the *j*th *n*-gram component C_j is given by:

$$w_j = g_j\left(\frac{1}{P(C_j|All)}\right), \tag{1}$$

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 classes (i.e. the prior probability of C_j). The components which have low 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 different classes.

It has been shown previously (for example, Zissman, 1996) that 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. In our phonotactic systems, we used three existing phone recognizers (Czech, Hungarian and Russian) from a toolkit developed by Brno University of Technology.[4] These were trained on the SpeechDat-E databases using a hybrid approach based on Neural Networks and Viterbi decoding (Matějka et al., 2005).

In addition, since our target applications are for English, we built an English decision-tree triphone-based phone recognizer, using the HMM toolkit (HTK) (Young et al., 2006). We trained the acoustic models using training data from the ABI-1 corpus (Section 2.2). The system uses 39 dimensional PLP-based feature vectors (Section 3.1). All phone HMMs comprise 3 emitting states without state-skipping, with one 16 component GMM per state. The phone recognizer uses a bigram phone-level language model derived from the ABI-1 training set. The pronunciation dictionary was generated from the British English Example Pronunciation dictionary (BEEP).[5]

For each phone recognizer and each class (language, accent or ethnic group), we built n-gram SVM language models for $n = 1$, 2, 3 and 4.

The result is our 'Phonotactic' system, obtained by fusing the outputs of all of the 16 individual phonotactic systems:

- *Phonotactics*: The outputs of our 16 phonotactic systems (English, Czech, Hungarian and Russian phone recognizers with unigram, bigram, trigram and 4-gram SVM language models) were fused using Brummer's multi-class linear logistic regression (LLR) toolkit.

In the accent and ethnic group recognition experiments, there is no development set to train the logistic regression fusing coefficients. Therefore we divided the test speakers (in each 'jackknife' round, in the case of accent recognition) into two sets. The class 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. The fused scores are then combined together and the final performance is calculated.

3.2.2. Acoustic systems

Modelling the distributions of acoustic features for different classes of speech has been successfully applied in language and speaker recognition systems. Most acoustic-based approaches use Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) for this purpose. The most common variants are GMM–UBM, GMM–SVM and GMM-n-gram (Torres-Carrasquillo et al., 2008), and these are described here.

- *GMM–UBM*: In the GMM–UBM approach, a Universal Background Model (UBM) is built using utterances from the training sets of all classes. Class-dependent models are obtained by MAP adaptation (Gauvain and Lee, 1994), adapting means and weights of the UBM, using the class-specific enrollment data. The result is one UBM and C class-dependent GMMs (where $C = 12$, 14 and 2 for language, accent and ethic group recognition, respectively). The inter-session variability within a class, such as inter-channel and inter-speaker variability, is estimated using the technique described in Vair et al. (2006). For a given test utterance, the component of the data that is attributable to inter-session variability can then be estimated, and the parameters of the UBM and class-dependent GMMs are adjusted to accommodate this variability.
- *GMM–SVM*: In our GMM–SVM system, the speech data from each individual speaker is 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 different classes are assumed to be linearly separable in this supervector space. The supervectors are used to build one SVM for each class, by treating that class as the 'target' and the others as the 'background' class.

  The connection between SVM scoring and GMM scoring is illustrated in Campbell and Karam (2009), where it is shown that the latter yields better results. Weighted averages of the target and background support vectors are 'pushed back' to the GMM domain and used in GMM scoring. Our system differs from that described in Campbell and Karam (2009) in two ways: first, only GMM means are adapted to form the supervectors, because this outperforms adapting the means and covariances. Second, inter-session compensation is applied to the target and non-target 'pushed' GMM models, as described for the GMM–UBM system above.

[4] http://www.fit.vubr.cz/research/groups/speech/sw/phnrec.

[5] "ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz" [cited April 30, 2011].

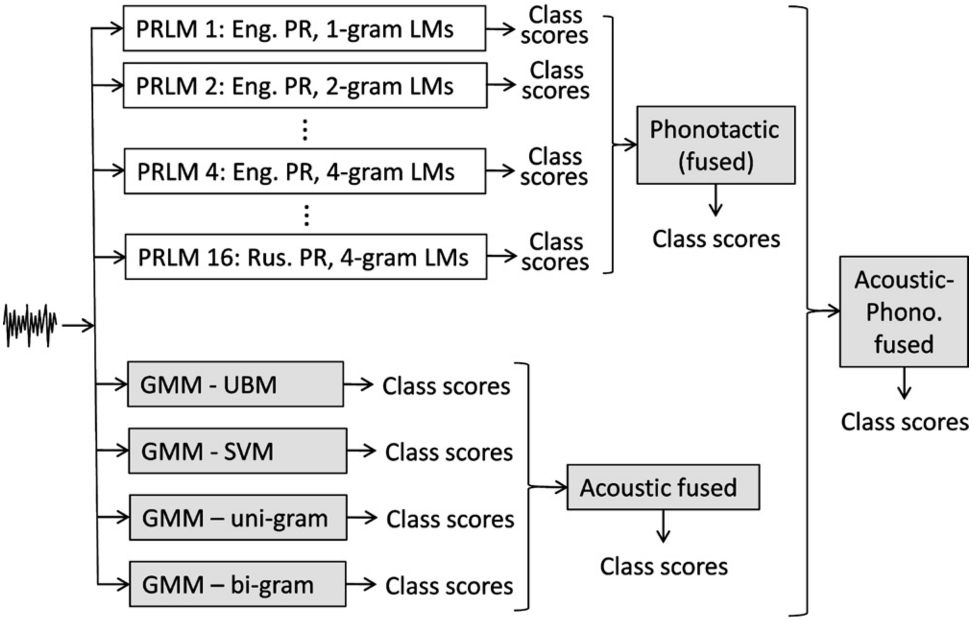

Fig. 2. Summary of text-independent phonotactic and acoustic LID systems (Eng. =,'English', Rus. = 'Russian', PR = 'Phone Recognizer'.

- *GMM-n-gram*: In the third acoustic based system, GMM-*n*-gram, the UBM GMM which is trained on the training data for all classes was used as a tokenizer (Fig. 1) to generate sequences of GMM component indices from the sequence of cepstral features. The resulting sequence is used to train 1-gram and 2-gram language models for each class using SVMs, as described in Section 3.2.1 and depicted in Fig. 1. Compared with the PRLM system described in Section 3.2.1, the phone recognizers are replaced by a class-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 component *n*-grams which represent the accent specific features and de-emphasize the components which represent features that are common in all accents (Hanani et al., 2010).
- *Acoustic fused*: 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.[6]
- *Acoustic–Phonotactic fused*: This system was obtained by fusing all four acoustic systems described in this section with the 16 phonotactics systems from the previous section, again using Brummer's toolkit.

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

$$S'_j = S_j - \underset{i \neq j}{max} S_i \tag{2}$$

Fig. 2 shows the components of the different text-independent phonotactic and acoustic LID systems that are described in Section 3.2. The systems depicted by shaded boxes correspond to the bullet-points in Section 3.2 and are evaluated in Section 6.

3.3. Text-dependent automatic systems

3.3.1. ACCDIST-based systems

In Wells (1982), British English accents are characterized according to differences in the realization of vowels in specific 'key words'. Huckvale's ACCDIST measure (Huckvale, 2007) makes this notion computationally useful.

[6] http://Niko.brummer.googlepages.com/focalmulticlass.

He argues that a relative measure, based on differences between the realizations of vowels in different words, is not only a cue for accent recognition, but is also less sensitive to other speaker specific characteristics than measures that depend on absolute spectral properties. Similar approaches are advocated in Barry et al. (1989) and Minematsu (2005). ACCDIST is text-dependent, since it requires a phone-level transcription of an utterance to which it is applied. Given a transcribed utterance, the start and end times of each realization of a vowel are identified. Each vowel segment is then split into two halves by time and the average feature vectors for each half (19 MFCCs plus energy) are concatenated to create a single 40 dimensional vector. The distances between these vectors, corresponding to different vowels in different contexts, are calculated using an unweighted Euclidean distance and stored in a distance table. To ensure that distance tables are comparable, all utterances must share the same phone-level transcription. An accent is represented as the average distance table over all of the training utterances for that accent. A test utterance is classified according to the correlation distance between its distance table and those for each of the accents.

In our variants of ACCDIST, a phonemic transcription of each of the SPA recordings in the ABI-1 corpus was generated using standard pronunciations from the BEEP dictionary. This was force-aligned with the speech data using our English phone recognizer (Section 3.2.1). Our ACCDIST-based system differs from that in Huckvale (2007) in two ways: first, we used the SPA data from ABI-1 rather than the “Short Sentence” files. The SPA recording 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 SPA recordings in our experiment, whereas only those recordings which were completed without errors or repetitions were used by Huckvale. This is because our ACCDIST systems do not require each recording to correspond to exactly the same phone sequence. Instead, the speaker distance tables are built from vowel tri-phone segments (i.e. “phone-vowel-phone”) rather than words. We also include vowel duration as an extra feature. For repeated tri-phones the mean feature vector was used. Hence, each SPA recording is represented as a sequence of pairs (v_i,p_i), where v_i is the 41-dimensional feature vector (two concatenated vectors comprising 19 MFCCs plus energy, plus vowel duration) of the *i*th tri-phone in the sequence and p_i is its label. The 105 most common tri-phones across all speakers in the training data were found and used in constructing the speaker distance tables. Our ACCDIST-based systems are as follows:

- *ACCDIST-Cor.dist.*: 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 performed using the correlation distance (Huckvale, 2007) between the test speaker distance table and the accent mean distance tables. The correlation took 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.
- *ACCDISI-SVM*: The success achieved by applying SVMs to supervectors constructed from stacked MAP-adapted GMM means in speaker and language recognition (Campbell et al., 2006) motivated us to apply SVMs to the speaker distance tables in our ACCDIST-based system. In our version of Huckvale's system above (ACCDIST-Cor.dist.), 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. Due to symmetry, each 105 × 105 distance matrix has 5460 distinct entries, which are rearranged into a 5460 dimensional vector. By labelling 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 vectorized distance table is evaluated against every accent model. The correlation distance kernel K was used in training and evaluating the SVM systems:

$$K(V_1, V_2) = \left[\frac{V_1 - \overline{V_1} I_J}{\sqrt{\sum_{j=1}^{J}(V_1^j - \overline{V_1})^2}}\right]\left[\frac{V_2 - \overline{V_2} I_J}{\sqrt{\sum_{j=1}^{J}(V_2^j - \overline{V_2})^2}}\right]^T \tag{3}$$

where J is the dimension of the vectorized distance tables (in this case $J = 5460$) V_1, V_2 are the two (5460 dimensional) distance table vectors, $\overline{V_1}$ and $\overline{V_2}$ are the (scalar) means of the coordinates of V_1 and V_2, respectively, and I_J is a J dimensional vector all of whose entries are equal to 1. $K(V_1, V_2)$ is the correlation distance kernel.

4. Human experiments

To provide baselines against which the automatic systems could be compared, two web-based human perceptual experiment were conducted.

- *Human regional accent recognition*: The human regional accent recognition experiment used exactly the same SPA test recordings as automatic classification. Twenty-four native British English speaking subjects, aged between 21 and 78, took part in the experiment. Each subject completed a registration process in which he or she gave their gender and age and indicated which, if any, of the 13 different ABI-1 locations they had ever lived in, and which of the 14 regional accents they were familiar with. Each subject then listened to a different set of 20 SPA recordings, each varying in length between 30 s and 40 s, selected randomly from the test set. For each recording, subjects were asked to identify the accent of the speaker (out of the 14 possible accents), the speaker's gender and age and to state their confidence in their decisions. The listeners were naïve in that they had no formal training in phonetics or linguistics and no explicit training in regional accent recognition was given. Instead the listeners were required to accomplish the task using the knowledge of regional accents that they had acquired naturally through their experiences to date.
- *Human ethnic group recognition*: The human ethnic group recognition experiment used exactly the same 315 test utterances that were used for automatic classification. Eight listeners who were familiar with the Birmingham accent took part in the experiment. As with the accent recognition experiment, the listeners had no formal background in phonetics or linguistics and no explicit training was given. Two subjects listened to all of the 315 test utterances, and six subjects listened to sets of 20 utterances. For each utterance, subjects were asked to identify the social group (Asian or White), to indicate their confidence in their decision, to estimate the age of the speaker, and to indicate the factors (acoustic quality, use of particular words or phrases, intonation, grammar, or other factors) that influenced their decision.

5. Experimental procedure

5.1. Phonotactic system training

For each phone recognizer (English, Czech, Hungarian and Russian) and each $n = 1, 2, 3, 4$, an n-gram SVM language model was trained using weighted n-gram probability vectors computed from the recognized phone sequences for each utterance in the training set (or, in the case of accent recognition, the training subsets of each 'jackknife' set).

5.2. Acoustic model training

Based on our experience with language identification (Hanani et al., 2010), unless otherwise stated all GMMs have 4096 components. UBM GMMs were created for language, accent and ethnic group recognition, using the corresponding training data with five iterations of the E-M algorithm. We used gender dependent models for LID and ethnic group recognition, and gender independent models for regional accent recognition. All of the GMM parameters (means, diagonal covariances and weights) were updated. Class-specific GMMs were MAP-adapted (Gauvain and Lee, 1994) from the corresponding UBMs, giving 12 language GMMs, 14 accent GMMs and two ethnic group GMMs, using the language-, accent- and ethnic-group-specific training material described in Section 2. The language, accent and ethnic group UBM GMM means were also MAP adapted using speech data from each speaker for each language, accent and social group, generating the GMM supervectors which are used to train the GMM–SVM systems and the projections for channel compensation (Section 3.2).

The same UBM GMMs (4096 components) were used as 'GMM tokenizers' to produce sequences of GMM component indices for the GMM-n-gram systems (Section 3.2).

All of the SVM models in this paper were trained and evaluated using the SVM-KM SVM MATLAB toolbox (Canu et al., 2005). The acoustic experiments were run on a PC incorporating an Nvidia Geforce GTX260 Graphics Processing Unit (GPU). Programming was done in MATLAB, GPUmat[7] and CUDA (NVIDIA, 2007).

[7] "http://gp-you.org", Accessed April 2011.

5.3. Recognition

For each system and each application we conducted verification and recognition experiments. Verification tests the claim that a particular utterance belongs to a particular class (language, regional accent, ethnic group) by comparing its score with a threshold. If the score exceeds the threshold then the claim is accepted. The Equal Error Rate (EER) is the error rate at the threshold where the False Acceptance Rate and False Rejection Rate are equal. We use the NIST LRE software to calculate percentage EER. Recognition simply assigns an utterance to the class with the largest score. The percentage accuracy (error) is the percentage of times that this class is correct (incorrect).

In the absence of transcriptions it is not possible to apply the ACCDIST-based systems to either language or social group identification.

6. Results and discussion

The results are summarized in Table 3. For each experiment the first entry is the percentage EER and the second entry is recognition accuracy.

6.1. Language identification

The performance of our text-independent LID system on the NIST 2003 evaluation set is reported elsewhere (Hanani et al., 2010). The figures are included here to establish that the system achieves state-of-the-art performance, to calibrate its various components, and to act as a reference point for the other experiments. The LID results for the different components of our system are shown in columns 2 (EER) and 3 (accuracy). The complete system (Acoustic–Phonotactic-fused) achieves 0.7% EER (98.42% language recognition accuracy). Of the four ‘acoustic’ systems, the best performance is obtained with the GMM–SVM system (0.92% EER), followed by GMM-unigram (2.82% EER). Fusing all of the acoustic systems gives an EER of 0.83%. The performance of the PRLM system (phonotactics, 1.48% EER) is better than each of the individual acoustic system except for GMM–SVM. Fusing all of the systems (Acoustics–Phonotactics-fused) gives the best result. We believe that the best performance on the NIST 2003 evaluation set published in 2003 was 2.8% EER (Singer et al., 2003) and that the overall best performance since then is 0.8% EER (Burget et al., 2006). This is comparable with our own best result (0.7% EER) and establishes the credibility of our system.

6.2. Accent identification

Columns four and five, and six and seven of Table 3 show accent recognition performance on the 30 s extracts and the SPA recordings, respectively. We focus on the SPA results (columns five and six), as these are available for all of the systems. The performance of each of the acoustic systems is poor despite the facts that a GMM with a large number of mixture components (4096) has been used and the recordings are good quality rather than telephone quality speech. To check that this is not due to the use of too many GMM components for the available training data, we repeated the GMM–UBM experiment using GMMs with 2^N components ($N = 4, \ldots, 12$). This confirmed that best performance is obtained with 4096 components.

The best acoustic performance, 8.3% EER, is achieved by fusing all of the acoustic systems. This indicates that at the acoustic level accent identification is a much more difficult task than speaker identification, where good performance on similar data can be achieved using a small number of GMM mixture components (Reynolds, 1995). It seems that for this type of data inter-speaker differences are the major source of variability, which masks inter-accent differences, and that the differences between the inventories of sounds in different accents in the same language are less pronounced than those between different languages. The PRLM system (phonotactics, 6.5% EER) outperforms all of the acoustic systems, despite the fact that phonotactic differences are apparently restricted because all of the recordings correspond to readings of the same text. The best accent recognition performance achieved with our language identification system is 4.52% EER (89.6% recognition accuracy), which is obtained by fusing all of the acoustic and phonotactic sub-systems.

The use of non-English (Czech, Hungarian and Russian) phone recognizers in the PPRLM system for English regional accent recognition requires some explanation. Table 2 shows the performances of the individual *n*-gram phonotactic systems ($n = 2, 3, 4$) for each phone recognizer (the unigram systems ($n = 1$) performed very poorly, and

Table 2
Performance EER [%] (*Accuracy [%]*) of phonotactic accent recognizers using English, Czech, Hungarian and Russian phone recognizers.

	English		Czech		Hungarian		Russian		PPRLM	
	30 s	SPA	30 s	SPA	30 s	SPA	30 s	SPA	30 s	SPA
2-Gram	25.2	22.7	26.8	26.4	23.4	23.4	28.6	22.6	17.7	16.8
	38	*38.6*	*39*	*37*	*46*	*43*	*32*	*48.2*	*52.5*	*55*
3-Gram	19.7	18.3	20.2	14.7	17	1.74	24.7	14.9	11.4	8.5
	50	*53*	*50*	*62.7*	*60.3*	*65*	*40*	*62*	*70*	*68.4*
4-Gram	18.4	14.8	17.7	9.6	14.2	10.8	23.9	14.6	9.8	6.7
	51	*58*	*55*	*68.7*	*62*	*71*	*42*	*63*	*73*	*79.6*
Fused	16.7	12.8	17.3	8.7	13.3	9.3	22.2	11.3	9.2	6.5
	56	*61*	*57*	*70.5*	*66*	*73*	*43*	*69*	*74*	*82*

Table 3
Summary of results for all systems and tasks. The figures are percentage Equal Error Rate (EER) and percentage recognition accuracy (Acc).

System	LID (NIST2003)		Accent ID (ABI-1) (30 s)		Accent ID (ABI-1) (SPA)		Ethnic group ID (VaB)	
	EER	Acc	EER	Acc	EER	Acc	EER	Acc
GMM–UBM (4096)	3.26	91.5	16.16	56.11	13.46	61.13	13.33	86.98
GMM–SVM (4096)	0.92	96.83	13.0	67.72	9.41	76.11	16.82	86.34
GMM-uni-gram	2.82	91.66	14.95	60.12	13.54	72.28	15.1	85
GMM-bi-gram	3.51	90.91	19.69	52.12	18.5	57.83	15.74	84.1
Acoustic-fused	0.83	97.3	12.33	73.6	8.3	77.32	7.28	92.7
Phonotactics	1.48	95.83	9.18	74.05	6.5	82.14	13.43	86.67
Acoustic–Phonotactic-fused	0.7	98.42	6.4	88.8	4.52	89.6	3.57	96.51
ACCDIST-Cor.dist.	–	–	–	–	2.66	93.17	–	–
ACCDIST-SVM	–	–	–	–	1.87	95.18	–	–
Human	–	–	–	–	–	58.24	–	90.24

their inclusion did not improve the overall performance of the fused systems). Focussing on recognition accuracy and the SPA test data, although the fused result is best with the Hungarian phone recognizer (73%), no individual system outperforms all of the others consistently and the performance obtained with the English phone recognizer is relatively poor. Ultimately, it seems that what is important is consistency rather than phone recognition accuracy. Referring to Fig. 1, in this application it may be better to regard these systems as abstract ‘tokenizers’ rather than explicitly as phone recognizers.

Both variants of the ACCDIST measure give better results than the acoustic-phonotactic LID system. The ACCDIST system with correlation distance gives 2.66% EER (93.17% accent classification accuracy), and the ACCDIST-SVM system gives the overall best result of 1.87% EER (95.18% accuracy). This compares with 92.3% accent recognition accuracy reported in Huckvale (2007). We conclude that exploiting linguistic knowledge about how the realization of vowels in particular contexts is indicative of regional accents of British English, gives a significant advantage compared to the purely data-driven approach that is followed in contemporary LID.

Human performance on the accent identification task (58.24% recognition accuracy) is significantly poorer than any of the automatic systems. However, this has also been observed in other studies (Arslan and Hansen, 1996). Table 4 shows the confusion matrix for the human accent recognition experiment, which is largely as one would predict. For example, there is confusion between the two northern English accents, East Yorkshire (ey) and Lancashire (la), the two Scottish accents, Glasgow (gl) and Scottish Highlands (sh), and between the two Irish accents, Dublin (ri) and Ulster (ul). The consistent misclassification of the North Wales accent (nw) as Liverpool (lp) is explained by the close geographical proximity of Denbigh, the town where the North Wales recordings were made, and Liverpool.

As expected, subjects are better at classifying accents with which they are familiar (Arslan and Hansen, 1996; Ikeno and Hansen, 2006). Human accent recognition accuracy is 76.2% for accents from regions where the listener has lived, compared with 51.7% for accents from regions where they have not lived, and 71.63% for ‘familiar’ accents and only 40.2% for ‘unfamiliar’ accents (according to the listeners’ responses to the questionnaire). The good performance

Table 4
Confusion matrix for the human regional accent recognition experiment (a sample corresponds to one listener classifying one utterance). Key: bm (Birmingham), cn (Truro, Cornwall), ea (Lowestoft, East Anglia), ey (Hull, East Yorkshire), gl (Glasgow), il (Inner-London), la (Burnley, Lancashire), lv (Liverpool), nc (Newcastle), nw (Denbigh, North Wales), ri (Dublin), sh (Elgin, Scottish Highlands), sse (Standard Southern English), ul (Belfast).

	bm	cn	ea	ey	gl	il	la	lv	nc	nw	ri	sh	sse	ul	Acc
bm	30	·	2	2	·	1	1	1	·	·	·	1	·	1	76·9
cn	·	15	6	·	·	1	1	·	1	3	1	·	2	1	48·4
ea	·	8	12	·	·	9	2	·	·	2	·	·	4	1	31·6
ey	·	·	6	19	1	1	14	·	3	·	·	·	1	·	42·2
gl	·	·	·	2	20	·	·	·	1	·	1	5	2	1	62·5
il	2	2	1	1	·	24	·	1	·	·	·	·	3	·	70·6
la	1	·	2	7	·	·	22	1	2	·	1	1	·	·	59·5
lv	·	·	·	·	·	·	·	28	3	·	·	·	2	·	84·9
nc	·	3	·	2	2	·	3	·	21	·	·	1	·	·	65·6
nw	·	·	2	4	·	·	3	11	2	10	·	·	3	1	27·8
ri	·	·	·	·	·	·	·	·	·	1	22	·	1	6	73·3
sh	2	·	·	1	9	·	·	·	·	1	·	19	·	1	57·6
sse	·	·	4	·	·	3	1	1	·	1	·	·	17	1	60·7
ul	·	·	1	·	1	·	·	·	·	·	8	1	·	20	64·5

Table 5
Confusion matrix for GMM–UBM regional accent recognition experiment (Key as for Table 4).

	bm	cn	ea	ey	gl	il	la	lv	nc	nw	ri	sh	sse	ul	Acc
bm	7	2	1	1	1	2	·	·	·	3	·	·	1	·	38·9
cn	·	7	·	·	·	7	·	1	·	1	·	2	·	·	38·9
ea	·	·	10	2	·	3	·	·	·	·	·	·	3	·	55·6
ey	·	1	·	11	·	·	3	·	1	2	·	·	·	·	61·1
gl	·	1	·	·	12	·	·	·	·	·	·	5	·	·	66·7
il	·	2	·	·	·	12	1	·	·	2	·	1	·	·	66·7
la	·	·	·	3	·	1	13	·	·	1	·	·	·	·	72·2
lv	·	·	·	2	·	2	·	11	·	3	·	·	·	·	61·1
nc	·	1	·	2	1	2	·	·	11	·	·	1	·	·	61·1
nw	·	·	·	1	·	3	1	1	·	12	·	·	·	·	66·7
ri	·	·	·	·	·	·	·	·	·	·	14	·	·	3	82·4
sh	·	·	·	·	·	1	·	·	·	·	·	17	·	·	94·4
sse	3	2	·	2	·	2	·	·	·	·	·	·	6	·	40·0
ul	·	·	·	1	·	2	·	·	·	1	4	1	·	8	47·1

for the Birmingham accent, and the overall shape of the confusion matrix, may be influenced by the presence of a disproportionate number of subjects from the Birmingham area in the listener group.

For comparison, Table 5 shows the corresponding confusion matrix for the acoustic GMM–UBM system, which achieves a similar accent recognition accuracy to the human listeners. The confusions are generally less intuitive. For example, as well as the expected confusion between Lancashire (la) and East Yorkshire (ey), there are many other instances of data being incorrectly classified as East Yorkshire or Inner-London, and examples where the Birmingham accent (bm) is incorrectly recognized as seven different accents.

6.3. Ethnic group identification

The final two columns of Table 3 show the results of the ethnic group identification experiments (Hanani et al., 2011). All of the acoustic systems achieve a similar level of performance, with the GMM–UBM giving the lowest EER at 13.33%. Fusion of the acoustic systems (‘Acoustic-fused’) results in an EER of 7.28%, a reduction of approximately 50% in EER relative to individual acoustic systems. This indicates that there is some orthogonality between the different acoustic systems for the ethnic group classification task. Despite the fact that the GMM–SVM system involves discriminative training it performs worse than the GMM–UBM system on this task. The PPRLM system

('Phonotactics') achieves similar performance to the best of the acoustic systems (13.43% EER). The final fused system ('Acoustic–Phonotactic-fused') scores 3.57% EER (96.51% accuracy). This compares with 9.76% EER (90.24% accuracy) for the human listeners.

The fact that the acoustic and phonotactic components contribute approximately equally to automatic social group identification performance is interesting. Subjectively, it is evident from listening to the recordings that the speech quality is different for the two groups, and one would expect this to be exploited by the acoustic systems. However it is also evident that the Asian recordings are characterized by the more frequent use of particular English words and the almost exclusive use of some non-English words (for example, people's names), which one would expect to be exploited by the phonotactic system. However, it seems that both of these phenomena contribute approximately equally to automatic classification performance. In contrast, the human listeners reported that in 75.5% of the tests the 'quality' of the speech contributed to their judgement, compared with 28% for the occurrence of specific words or phrases, 23.8% for intonation, 11.8% for grammar, and 0.6% for 'other factors'.

7. Conclusions

The objective of this paper is to measure the ability of a state-of-the-art automatic LID system to extract two particular types of paralinguistic information from a speech signal, specifically the regional accent of the speaker, and the ethnic group to which he or she belongs. In both cases, the performance of the LID system is compared with human performance, and for regional accent recognition, with other automatic systems based on Huckvale's ACCDIST measure. The "Accents of the British Isles" (ABI-1) corpus of good quality recordings of read speech, representing 14 different regional accents of spoken British English, is used for our experiments in accent recognition. The "Voices across Birmingham" (VaB) corpus of telephone conversational speech between subjects who were born and live in the city of Birmingham (UK) is used for ethnic group recognition.

For both regional accent and ethnic group recognition, automatic LID outperforms naïve human listeners. The recognition error rate for human listeners is approximately four times greater than that for the LID system for regional accent recognition (41.76% compared with 10.4%), and three times greater for ethnic group recognition (9.76% compared with 3.49%).

Regional accent recognition appears to be a challenging task for both automatic systems and human listeners. Even though the ABI-1 recordings are good quality read speech (rather than telephone conversational speech), the best performance of our LID system on 30 s segments is 6.4% EER (88.8% accuracy) compared with 0.7% EER (98.42% accuracy) for LID using the same amount of telephone conversational speech from the NIST 2003 evaluation. The best regional accent recognition performance is 1.87% EER (95.18% accuracy), which is achieved using the ACCDIST-SVM system and the SPA recordings. The superior performance of the ACCDIST-based systems relative to the LID system is an interesting example where the explicit use of linguistic knowledge results in a method that outperforms a purely data-driven statistical approach, and with a much lower computational requirement. However, a clear disadvantage of the ACCDIST method is its text dependency, in that transcriptions of the training and test utterances are required and they must be comparable in terms of the vowel contexts that they include. A clear future opportunity is to exploit the ideas that motivate ACCDIST without relying on a such a transcription.

Regionally accented speech in the ABI-1 corpus is defined to be speech spoken by an individual who was born in that region and has lived there for all of his or her life. However, even with this residency constraint many subjects' accents exhibit non-regional influences. It seems that naïve native human listeners can correctly place characteristic examples of regionally accented speech but have difficulty in cases where, for example, as a consequence of social or educational factors a subject's accent exhibits strong traits of Standard English. However, the relatively good performances of the automatic systems indicate that correct classification of many of these more subtle instantiations of regional accent is possible. It would be interesting to know how well human listeners can perform given suitable explicit training.

Intuitively, the ethnic group classification task appears to be more difficult than accent recognition, even though it is a two-class problem, since the classes share some aspects of the same regional accent, and the data is telephone conversational speech. However, the acoustic and phonotactic components of our automatic LID system score recognition accuracies of 92.7% and 86.7%, respectively, and the overall best performance is 96.5% accuracy, achieved by fusing all of the acoustic and phonotactic subsystems. This result is much better than expected and compares with an accuracy of 90.24% for human listeners. As in the case of regional accent recognition, it would be interesting to know how well human listeners would perform if they were given explicit training for this task.

The fact that it is possible to access these types of paralinguistic information using as little as 30 s of data has interesting implications for automatic speech recognition. It confirms that there are significant acoustic and phonotactic differences between, and even within, regional accents, and it shows that these differences are sufficiently large be detected automatically. Hence it may be possible to use these technologies to identify suitable acoustic, lexical and even grammatical models automatically, as a first step towards rapid adaptation.

Although it is difficult to make direct comparisons, it seems that our LID system performs better on the language recognition and the ethnic group recognition tasks than on the regional accent recognition task, even though the former are based on conversational speech recorded over a telephone channel while the latter involves good quality recordings of read speech. It would be interesting to know whether the availability of natural conversational speech is advantageous for these types of paralinguistic tasks.

Finally, and returning to our original premise, we conclude from the results presented in this paper that the distributions of acoustic feature vectors, and phone *n*-grams, corresponding to different regional accents of English or different ethnic groups within an accent, are sufficiently distinct to enable pattern recognition methods from LID to be applied successfully to automatic regional accent and ethnic group classification. From a broader perspective, this raises the possibility of applying similar techniques to the automatic classification of other paralinguistic phenomena.

References

Angkititrakul, P., Hansen, J.H.L., 2006. Advances in phone-based modeling for automatic accent classification. IEEE Transactions on Audio, Speech and Language Processing 14 (2), 634–646.

Arslan, L.M., Hansen, J.H.L., 1996. Language accent classification in American English. Speech Communication 18, 353–367.

Barry, W.J., Hoequist, C.E., Nolan, F.J., 1989. An approach to the problem of regional accent in automatic speech recognition. Computer Speech and Language 3, 355–366.

Biadsy, F., Hirschberg, J., Habash, N., 2009. Spoken Arabic dialect identification using phonotactic modeling. In: Proc. EACL 2009 Workshop on Computational Approaches to Semitic Languages, Athens, Greece.

Biadsy, F., Hirschberg, J., Collins, M., 2010. Dialect Recognition using Phone-GMM-Supervector-based SVM Kernel. In: Proc. InterSpeech 2010, Makuhari, Chiba, Japan, pp. 753–756.

Burget, L., Matějka, P., Černocký, J., 2006. Discriminative training techniques for acoustic language identification. In: Proc. IEEE ICASSP 2006, pp. I-209–I-212.

Campbell, W.M., Campbell, J.P., Reynolds, Jones, D.A., Leek, T.R., 2004. Phonetic speaker recognition with support vector machines. Advances in Neural Information Processing Systems 16.

Campbell, W.M., Sturim, D.E., Reynolds, D.A., 2006. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters 13 (5), 308–311.

Campbell, W.M., Karam, Z.N., 2009. A Framework for discriminative SVM/GMM systems for language recognition. In: Proc. Interspeech 2009, Brighton, UK, pp. 2195–2198.

Canu, S., Grandvalet, Y., Guigue, V., Rakotomamonjy, A., 2005. SVM and Kernel Methods Matlab Toolbox. Perception Systèmes et Information, INSA de Rouen, Rouen, France, http://asi.insa-rouen.fr/enseignants/arakotom/toolbox/.

D'Arcy, S.M., Russell, M.J., Browning, S.R., Tomlinson, M.J., 2005. The accents of the British Isles (ABI) corpus. In: Proc. Modélisations pour l'Identification des Langues, MIDL Paris, pp. 115–119.

D'Arcy, S.M., 2007. The effect of age and accent on automatic speech recognition performance. Ph.D. Thesis, University of Birmingham, Birmingham, UK.

Elmes, S., 2005. Talking for Britain: A Journey through the Nation's Dialects. Penguin Books.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verificatio. IEEE Transactions on Acoustics Speech and Signal Processing 29, 254–272.

Gauvain, J.-L., Lee, C., 1994. Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing 2, 291–298.

Hanani, A., Carey, M., Russell, M., 2010. Improved language recognition using mixture component statistics. In: Proc. Interspeech 2010, Makuhari, Chiba, Japan, pp. 741–744.

Hanani, A., Russell, M., Carey, M., 2011. Speech-based identification of social groups in a single accent of British English by humans and computers. In: Proc. IEEE ICASSP 2011, pp. 4876–4879.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing 2 (October (4)), 578–589.

Huang, R., Hansen, J.H.L., Angkititrakul, P., 2007. Dialect/accent classification using unrestricted audio. IEEE Transactions on Audio, Speech and Language Processing 15 (2), 453–464.

Huckvale, M., 2007. ACCDIST: an accent similarity metric for accent recognition and diagnosis. In: Müller, C. (Ed.), Speaker Classification II. Springer-Verlag, Berlin/Heidelberg, Germany, pp. 258–275.

Hughes, A., Trudgill, P., Watt, D., 2005. English Accents and Dialects: An Introduction to the Social and Regional Varieties of English in the British Isles, fourth ed. Hodder Arnold.

Humphries, J.J., Woodland, P.C., 1997. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. In: Proc. 5th European Conference on Speech Communication and Technology, Eurospeech'97, pp. 2367–2370.

Ikeno, A., Hansen, J.H.L., 2006. Perceptual recognition cues in native English accent variation: 'Listener accent, perceived accent, and comprehension'. In: Proc. IEEE ICASSP 2006, vol. I, pp. 401–404.

Lincoln, M., Cox, S., Ringland, S., 1998. A comparison of two unsupervised approaches to accent identification. In: Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP 98), Sydney, Australia.

Matějka, P., Schwarz, P., Černocký, J., Chytil, P., 2005. Phonotactic language identification using high quality phoneme recognition. In: Proc. Interspeech'2005, Lisbon, pp. 2237–2240.

Miller, D.R., Trischitta, J., 1996. Statistical dialect classification based on mean phonetic features. In: Proc. 4th Int. Conf. on Spoke Language Processing (ICSLP 96), Philadelphia, PA, USA.

Minematsu, N., 2005. Mathematical evidence of the acoustic universal structure in speech. In: Proc. IEEE ICASSP 2005, pp. 889–992.

NVIDIA, 2007. NVIDIA CUDA Compute Unified Device Architecture: Programming Guide.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. In: Proc. Odyssey 2001, Speaker and Language Recognition Workshop, Crete, Greece, pp. 213–218.

Purnell, T., Idsardi, W., Baugh, J., 1999. Perceptual and phonetic experiments on American English dialect identification. Journal of Language and Social Psychology 18 (1), 10–30.

Reynolds, D.A., 1995. Large population speaker identification using clean and telephone speech. IEEE Signal Processing Letters 2, 46–48.

Richardson, F.S., Campbell, W.M., Torres-Carrasquillo, P.A., 2009. Discriminative *N*-gram selection for dialect recognition. In: Proc. Interspeech 2009, Brighton, UK.

The SCRIBE Manual, 1998. http://www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm.

Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A., 2003. Acoustic, phonetic, and discriminative approaches to automatic language identification. In: Proc. Eurospeech 2003, pp. 1345–1348.

Teixeira, C., Trancoso, I., Serralheiro, A., 1997. Recognition of non-native accents. In: Proc. 5th European Conference on Speech Communication and Technology, EUROSPEECH'97, pp. 2375–2378.

Tjalve, M., Huckvale, M., 2005. Pronunciation variation modelling using accent features. In: Proc. 9th European Conference on Speech Communication and Technology, Interspeech 2005, Lisbon, Portugal.

Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Deller Jr., J.R., 2002. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In: Proc. Int. Conf. Spoken Language Processing, ICSLP 02, Denver, USA, pp. 89–92.

Torres-Carrasquillo, P.A., Singer, E., Campbell, W.M., Gleason, T., McCree, A., Reynolds, D.A., Richardson, F., Shen, W., Sturim, D., 2008. The MITLL NIST LRE 2007 language recognition system. In: Interspeech'08, Brisbane, Australia.

Vair, C., Colibro, D., Castaldo, F., Dalmasso, E., Laface, P., 2006. Channel factors compensation in model and feature domain for speaker recognition. In: Proc. IEEE Speaker and Language Recognition Workshop, Odyssey 06.

Walton, J.H., Orlikoff, R.F., 1994. Speaker race identification from acoustic cues in the vocal signal. Journal of Speech and Hearing Research 37, 738–745.

Wells, J.C., 1982. Accents of English, Volume 1: An introduction. Cambridge University Press.

Wells, J.C., 1982. Accents of English, Volume 2: The British Isles. Cambridge University Press.

Woehrling, C., Boula de Mareüil, P., 2006. Identification of regional accents in French: perception and categorization. In: Proc. Interspeech 2006, Pittsburgh, PA, USA, pp. 1511–1514.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Guan, Y., Hu, R., Oura, K., Wu, Y.-J., Tokuda, K., Karhila, R., Kurimo, M., 2010. Thousands of voices for HMM-based speech synthesis – analysis and application of TTS systems built on various ASR corpora. IEEE Transactions on Audio, Speech and Language Processing 18 (5), 984–1004.

Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2006. The HTK Book Version 3.4. Cambridge University Press.

Zhai, L.F., Siu, M., Yang, X., Gish, H., 2006. Discriminatively trained language models using support vector machines for language identification. Proc. IEEE Speaker and Language Recognition Workshop, Odyssey 06, 1–6.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Speech and Audio Processing 4 (1), 31–44.

Zissman, M.A., Gleason, T.P., Rekart, D.M., Losiewicz, B.L., 1996. Automatic dialect identification of extemporaneous, conversational, Latin American Spanish speech. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc, ICASSP'96, vol. 2, pp. 777–780.