SPEECH-BASED IDENTIFICATION OF SOCIAL GROUPS IN A SINGLE ACCENT OF BRITISH ENGLISH BY HUMANS AND COMPUTERS

Abualsoud Hanani, Martin Russell and Michael J. Carey

School of Electronic, Electrical and Computer Engineering, University of Birmingham
{aah648, m.j.russell, m.carey}@bham.ac.uk

ABSTRACT

Classification of social groups within a given accent is a challenging refinement of language identification (LID) and accent/dialect recognition. The 2001 census of England and Wales identifies two main ethnic groups in the city of Birmingham, which it refers to as Asian and white. In this paper LID techniques are applied to the problem of identifying individuals from these two groups who were born in Birmingham and hence speak British English with a Birmingham accent. An Equal Error Rate (EER) of 3.57% is obtained using a LID system which fuses the outputs of several acoustic and phonotactic systems. This performance is much better than expected and compares to an EER of 8.72% achieved by human listeners. The implications of this result for automatic speech recognition are discussed.

Index Terms— Language identification, accent recognition, dialect recognition

1. INTRODUCTION

Spoken British English can be partitioned into a range of regional accents and dialects [1]. However, even within a particular accent region there is variation – for example, people born and raised in different neighborhoods or in different social groups in the same city can often be distinguished by their speech. From the perspective of speech technology, these systematic differences in spoken language are important because they may offer an approach to fast characterization of new talkers.

Researchers engaged in automatic accent or dialect recognition (e.g. [2]) have tended to adopt similar techniques to those used in Language Identification (LID). The purpose of this paper is to determine whether these same techniques can distinguish between different groups within the same accent.

Various cues used by humans and machines to distinguish between languages have been explored in previous research, including phone inventory, phonotactics, prosody, lexical cues, morphology, and syntax. Some of the most successful approaches to LID are those based on phonotactic variation. A typical phonotactic-based system is described in the classic paper by Zissman [3]. In the Phone Recognition-Language Modelling (PRLM) approach, a Phone Recognizer (not necessarily trained on a related language) is first used to estimate the phone sequence for an utterance, and a set of Language Models is used to estimate the probability that this phone sequence was spoken in a particular language.

Similar approaches have been applied to accent identification. For example, Zissman et al. [4] used the PRLM approach to distinguish between Cuban and Pervian dialects of Spanish, with an English phone recognizer trained on TIMIT data. The accuracy of this system is 84%. In another study, Biadsy et al [5] used the same approach to classify five Arabic dialects, using eight parallel phone recognizers trained on different languages. This system achieved an accuracy of 81.6% on the five dialects. Torres-Carrasquillo et al [6] studied an alternative approach to identifying Cuban and Pervian Spanish, using Gaussian Mixture Models (GMM) with shifted-delta-cepstral acoustic features. This performs less accurately (70%) than the phonotactic system in [4]. Finally, Richardson et al [7] achieved state-of-the-art dialect recognition by fusing the phonotactic and acoustic systems.

In this paper we consider the problem of identifying individuals from two social groups who speak English with the same regional accent, namely Asian and white people who were born and live in Birmingham, UK. An Equal Error Rate (EER) of 3.57% is obtained using a LID system which fuses the outputs of several acoustic and phonotactic systems. This performance is better than expected and compares to an EER of 8.72% achieved by human listeners.

2. SPEECH DATA

The goal of the "Voices across Birmingham (VaB)[1]" project is to capture variations in conversational speech across the people of the city of Birmingham in the UK. It currently comprises around 175 hours of recordings of telephone conversational speech between participants who were born in or around the city. Each participant made up to one hour 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.

[1] http://www.thespeechark.com

The 2001 census of England and Wales[2] included questions about the ethnicity of residents. According to the results, at that time approximately 70% of Birmingham's population categorized themselves as 'white' and 20% as 'Asian'. The VaB project asked its participants similar questions, and for these two 'majority' groups there is sufficient data to conduct an experiment to study whether or not an individual can be classified automatically into the correct social group from his or her speech.

The British Asian group can be further sub-divided into those 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 seconds. 175 speakers are Asian (69 male, 106 female) and 140 are white (53 male, 87 female).

3. HUMAN PERFORMANCE

To provide a baseline against which the automatic social-group-recognition systems could be compared, a web-based human perceptual experiment was conducted using exactly the same 315 test utterances. Two subjects listened to all of the 315 test utterances, and a further 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, or other factors) that influenced their decision. The human listeners scored an average Equal Error Rate (EER) of 8.72% for the social-group identification task.

4. DESCRIPTION OF AUTOMATIC SYSTEMS

4.1 Phonotactic Systems

The success of the PRLM approach for language and dialect identification motivated us to apply it to our social group identification task. Recall that, in the PRLM approach a sequence of phones is extracted from each training utterance from the two social groups 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 group. Before building the language models, a weighting technique proposed in [8] and used in our language ID system in [9] is applied to the *n*-gram probabilities in order to emphasize the most discriminative components (i.e. those which are common in one group but not in the other). This weighting also de-emphasizes the *n*-gram components that are common in both groups, as they do not carry useful information for discrimination. The weight w_j for component C_j is given by:

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

Where g_j is a function used to smooth and compress the dynamic range (for example, $g_j(x)=\sqrt{x}$, or $g_j(x)=\log(x)+1$). $p(C_j/All)$ is the probability of *n*-gram component C_j across the two groups. The components which have zero occupancy in the two groups 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 SVM models for the two social groups.

Using multiple PRLM models with phone recognizers trained on different languages and combining them in the back end has been shown to improve the performance of language and dialect ID systems [3].

In our Phonotactic systems, we have used four different phone recognizers for English (Eng), Czech (Cze), Hungarian (Hun) and Russian (Rus), from a toolkit developed by Brno University of Technology[3]. The English phone recognizer was trained on TIMIT, while Czech, Hungarian and Russian phone recognizers were trained on the SpeechDat-E databases using a hybrid approach based on Neural Networks and Viterbi decoding.

4.2 Acoustic Systems

Modeling acoustic features extracted from speech is an alternative method that has been successfully used in language and dialect recognition systems. The Gaussian Mixture Model (GMM) is the core of most acoustic based approaches, of which GMM-UBM, GMM-SVM and GMM-n-gram are the most successful.

In the GMM-UBM system, two gender-dependent Universal Background Models (UBMs) were trained with the EM-algorithm, using utterances from both social groups. Social-group-dependent models are obtained by MAP adaptation (adapting means only) of the UBM, using the group-specific enrollment conversations. The result is two UBMs and four social-group- and gender-dependent models.

In our GMM-SVM system, each single utterance 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'. Hence each speech

[2] http://www.statistics.gov.uk/census2001

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

utterance is mapped from the cepstral feature vector sequence domain to the supervector domain, where the classes are assumed to be linearly separable. This process also normalizes the length of the utterances. The supervectors are used to build one SVM model for each social group, by taking one group as a 'target' class and the other as a 'background' class.

In the third acoustic based system, GMM-*n*-gram, the gender- and social-group-dependent GMMs are used as tokenizers to generate sequences of GMM component indeces from the sequence of cepstral features. The resulting sequences are used to train an *n*-gram language model for each social group, using SVMs. Compared with the phonotactic system (PRLM) described earlier, the phone recognizer is replaced by an acoustic GMM producing a sequence of indeces for the *n*-best Gaussian components instead of sequence of phones. The other parts of these two types of system are the same, including the use of the discriminative weighting technique to emphasize the GMM components which represent the social group specific features and de-emphasize the components which represents the common features in both groups.

In all our systems, the score of one group model is normalized with the other model.

4.3 Fusion

The outputs of *n*-grams (*n*=1,2,3 and 4) of the four phone recognizers were combined and fused with Brummer's 2-class (target and non-target) linear logistic regression (LLR) toolkit[4] (column 6 in Table1). (2-4)-grams of each single Phone Recognizer were also fused with LLR (row 6 in Table1). In addition, (2-4)-grams of the four Phone Recognizers (12 systems) were also fused together, giving the best performance of all of the Phonotactic systems.

In the same way, the outputs of the three acoustic-based systems were fused together (row 4 in Table2), and also fused with the best Phonotactic system.

Because there is no development set to train the logistic regression and find the best fusing coefficients, we divided the 315 testing utterances into two different sets; one with 157 utterances and the other with 158 utterances. The social group and gender of speakers are distributed equally in both sets. One set is used to find the coefficients to be used in fusing the systems on the second set, and vice versa. The fused scores are then combined together and the final performance is estimated. This method was used to obtain all of the fusion results in Tables 1 and 2.

4.4 Experimental Setup

Four different orders of n-grams, 1, 2, 3 and 4-gram, are used to model the phone sequences produced by the four phone recognizers described in Section 3.1 (Rows 2, 3, 4, and 5 in Table 1). This results in sixteen PRLM systems: four phone recognizers times four *n*-gram systems. For each *n*-gram, the four PRLM systems are combined together (PPRLM) with Linear Logistic Regression (LLR) (column 6 in Table1).

In the acoustic level experiments, acoustic 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 their deltas features, giving a total of 38 features per frame at a frame rate of 100 frame per second. RASTA filtration is applied to the power spectra and feature warping, with 3-seconds windows, is applied on the final feature vectors to reduce the channel effect.

Two gender dependent UBMs, each with 4094 mixture components, were trained on the acoustic training data with 5 EM iterations updating all parameters; means, diagonal covariances and weights. Four social-group dependent GMMs were MAP-adapted from the UBMs using group specific data (row 2 in Table 2). The UBM means were also MAP adapted using each single-side conversation of each group, generating the GMM supervectors which were used to train the GMM-SVM system (row 3 in Table 2).

Four 4096 component gender- and social-group-dependent GMMs are trained on the corresponding acoustic data, and then used as GMM-tokenizers to produce a sequence of GMM component indeces. A uni-gram system models the output of the four GMM tokenizers. The two social-group-dependent systems are then combined at the end with LLR (row 3 in Table 2) in the same way that parallel PRLM systems are combined in the phone based systems.

All *n*-gram systems and the GMM-SVM are modeled using the free SVM-KM[5] SVM MATLB toolbox. Computations in the acoustic experiments were accelerated by an Nvidia Geforce GTX260 Graphics Processing Unit (GPU), comprising 216 floating-point processors and 1.76GB RAM together with an Nvidia C106 Tesla machine with approximately similar performance. Programming was carried out in MATLAB, GPUmat and CUDA.

5. RESULTS AND DESCUSSION

The experimental results for the sixteen PRLM phonotactics systems, using English, Czech, Hungarian and Russian phone recognizers, are presented in Table 1. The results are presented as percentage EER on the 315 40-second test utterances.

As is clear from results in Table 1, combining parallel PRLM systems (PPRLM) with different phone recognizers improves the performance of all *n*-gram systems. Fusing the four *n*-gram systems for each single phone recognizer also

[4] www.dsp.sun.ac.za/~nbrummer/focal/

[5] http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox

improves the system performance. The best performance (13.23%EER) is obtained by fusing the 2, 3 and 4-gram systems for the four phone recognizers together. Given that the task is to discriminate between closely related variations of British English, the PRLM performance might be improved by the use of a more appropriate British English phone recognizer rather than the American English, Czech, Hungarian and Russian systems used in this experiment.

Table 1: The performance (EER %) of the Phonotactic based machine systems using different phone recognizers.

Phone Recog	Eng	Cze	Hun	Rus	Combined (PPRLM)
1-gram	44.34	44.6	38.14	43.02	31.83
2-gram	26.35	23.65	20.68	17.96	16.08
3-gram	23.79	19.84	18.69	20.31	18.35
4-gram	23.67	21.36	24.61	26.37	18.11
2,3,4-fused	21.55	15.61	18.35	18.25	13.23

The performance of the acoustic based machine systems are presented in Table 2. The best performance (13.33%) among the three acoustic systems is obtained by the GMM-UBM system (second row in Table 2). This performance is dramatically improved (45.39%) by fusing the GMM-UBM with the other two acoustic systems (row 5 in Table 2). A further 50.96% improvement is obtained when fusing the acoustic systems with the best Phonotactic system (row 6 in Table 2).

Table 2: Performance (EER%) of the acoustic based systems

Acoustic System	EER [%]
GMM-UBM	13.33
GMM-SVM	16.82
Combined GMM-uni-gram	15.08
Acoustics-Fused	7.28
Phonotactic-Acoustic-Fused	3.57

6. CONCLUSION

In this paper we investigated whether techniques used for language identification and accent/dialect recognition are able to distinguish between talkers from different social groups within a single regional accent.

The 2001 census of England and Wales identifies two main ethnic groups in the city of Birmingham, UK, namely Asian and white. These groups are well represented in "Voices across Birmingham", a corpus of recordings of telephone conversational speech between individuals in the city. In this study we only consider speech from those participants who were born in Birmingham.

The results of applying various acoustic and phonotactic LID systems to this problem are reported. The best phonotactic and acoustic systems score EERs of 13.23% and 7.28%, respectively. The overall best performance (3.57% EER) is achieved using a system which fuses the outputs of a combination of these acoustic and phonotactic systems. This result is much better than anticipated and compares with an EER of 8.72% for human listeners

The fact that it is possible to decide automatically which social group within a particular accent group an individual belongs to, and to achieve this using as little as 40s of data, has interesting implications for automatic speech recognition. First, it confirms that there are significant acoustic and phonotactic differences even within a 'homogeneous' accent group. Second, it shows that these differences are sufficiently large be detected automatically. Hence it may be possible to identify suitable acoustic, lexical and even grammatical models automatically for rapid adaptation.

7. REFERENCES

1. J. C. Wells, "Accents of English 2: The British Isles", *Cambridge University Press*, 1982.
2. L. Arslan and J.H.L. Hansen, "Language Accent Classification in American English". *Speech Communication*. **18**(4): p. 353-367, July 1996.
3. M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech". *IEEE Trans. on Speech and Audio Proc*., **4**(1): p. 31-44, 1996.
4. M. A. Zissman, T. P. Gleason, D. M. Rekart and B. L. Losiewicz, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP-96, 1996.*
5. F. Biadsy, J. Hirschberg and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling", in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*. 2009,
6. P. A. Torres-Carrasquillo, T.P. Gleason and D. A. Reynolds, "Dialect Identification Using Gaussian Mixture Models", *Proc. Odyssey: The Speaker and Language Recognition Workshop*, ISCA, p. 297-300, 2004.
7. F. S. Richardson, W.M. Campbell, and P.A. Torres-Carrasquillo, "Discriminative N-Gram Selection for Dialect Recognition", *Proc. Interspeech 2009,* Brighton, UK., 2009.
8. W. M. Campbell , J. P. Campbell , D. A. Reynolds , D. A. Jones , T. R. Leek, "Phonetic speaker recognition with support vector machines,". in *Advances in Neural Information Processing Systems* 16, 2004.
9. A. Hanani., M. Carey, and M. Russell, "Improved Language Recognition using Mixture Component Statistics", *Proc. Interspeech 2010*, Tokyo, Japan,, 26-30 September 2010.