Dialect Classification for online podcasts fusing Acoustic and Language based Structural and Semantic Information

Rahul Chitturi, John. H.L. Hansen[1]
Center for Robust Speech Systems(CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas
Richardson, Texas 75080, U.S.A
{rahul.ch@student, john.hansen@}utdallas.edu

Abstract

The variation in speech due to dialect is a factor which significantly impacts speech system performance. In this study, we investigate effective methods of combining acoustic and language information to take advantage of (i) speaker based *acoustic* traits as well as (ii) content based word selection across the *text* sequence. For acoustics, a GMM based system is employed and for text based dialect classification, we proposed n-gram language models combined with Latent Semantic Analysis (LSA) based dialect classifiers. The performance of the individual classifiers is established for the three dialect family case (DC rates vary from 69.1%-72.4%). The final combined system achieved a DC accuracy of 79.5% and significantly outperforms the baseline acoustic classifier with a relative improvement of 30%, confirming that an integrated dialect classification system is effective for American, British and Australian dialects.

1 Introduction

Automatic Dialect Classification has recently gained substantial interest in the speech processing community (Gray and Hansen, 2005; Hansen et al., 2004; NIST LRE 2005). Dialect classification systems have been employed to improve the performance for Automatic Speech Recognition (ASR) by employing dialect dependent acoustic and language models (Diakoloukas et al., 1997) and for Rich Indexing of Spoken Document Retrieval Systems(Gray and Hansen 2005). (Huang and Hansen, 2005; 2006) focused on identifying pronunciation differences for dialect classification. In this study, unsupervised MFCC based GMM classifiers are employed for pronunciation modeling. However, English dialects differ in many ways other than pronunciation like Word Selection and Grammar, which cannot be modeled using frame based GMM acoustic information. For example, word selection differences between UK and US dialects such as - "lorry" vs. "truck", "lift", vs. "elevator", etc. Australian English has its own lexical terms such as tucker (food), outback (wilderness), etc (John Laver, 1994). N-gram language models are employed to address these problems. One additional factor in which dialects differ is in Semantics. For example, *momentarily* which means for a *moments duration* (UK) vs. *in a minute or any minute now* (US). The sentence "*This flight will be leaving momentarily"* could represent different time duration in US vs. UK dialects (John Laver, 1994). Latent Semantic Analysis is a technique that can distinguish these differences (Landauer et al.,1998). LSA has been shown to be effective for NLP based problems but has yet to be applied for dialect classification. Therefore, we develop an approach that uses a combination with n-gram language modeling and LSA processing to achieve effective language based dialect classification accuracy. Sec 4 explains the baseline acoustic classifier. Language classifiers are described in Sec 5 and the results which are presented in Sec 6 affirm that combining various sources of information significantly outperforms the traditional (or individual) techniques used for dialect classification.

[1]This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029

2 Online Podcast Database

The speech community has no formal corpus of audio and text across dialects of common languages that could address the problems discussed in Sec.1. It was suggested in (Huang and Hansen, 2007) that it is more probable to observe semantic differences in the spontaneous text and speech rather than formal newspapers or prepared speeches since they must transcend dialects of a language (Hasegawa-Johnson and Levinson, 2006; Antoine 1996). Therefore, we collected a database from web based online podcasts of interviews where people talk spontaneously. All these are already been transcribed in order to separate text and audio structure and to temporarily set aside automatic speech recognition (ASR) error. These podcasts are not transcribed with an exact word to

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 21–24,
Columbus, Ohio, USA, June 2008.

word match but they match the audio to an extent that include what the speakers intended to say. The language and Acoustic statistics of this database are described in Sec 2.1, and 2.2.

2.1 Language Statistics

Huang and Hansen observed that the best dialect classification accuracy for N-gram classification requires at least 300 text words to obtain reasonable performance (Huang and Hansen, 2007). So, these interviews are segmented into blocks of text with an average text of 300 words. Table 1 summarizes the text material for three family-tree branches of English, containing 474k words and 1325 documents.

Dialect	No.of words	No. of Documents	
		Train	Test
US English	200k	383	158
UK English	154k	288	122
AU English	120k	233	141

Table 1: Language Statistics

2.2 Acoustic Statistics

We note that the data collected from online podcasts is not well structured. The audio data is segmented into smaller audio segment files since we are interested in 300 word blocks. Since the collection of dialect podcasts are collected from a wide range of online sources, we assume that channel effects and recording conditions are normalized across these three dialects. We also note that there is no speaker overlap between the test and train data. Therefore, there are no additional acoustic clues other than dialect. Table 2 summarizes the acoustic content of the corpus with 231 speakers and 13.5 hrs of audio.

Dialect	Males	Females	No. of Hours	
			Train	Test
US English	48	37	3.2	1.7
UK English	40	32	2.3	1
AU English	36	38	3.3	2

Table 2: Acoustic Statistics

3 System Architecture

The system architecture is shown in Fig 1, which consists of two main system phases for acoustic and language classifiers. MFCC based classifiers are used for acoustic modeling, while for language modeling, we use a combination of n-gram language modes and LSA classifiers. In the final phase, we combine the acoustic and language classifiers into our final dialect classifier. To construct the overall system, we first train the individual classifiers, and then set the weights of the hybrid classifiers using a greedy strategy to form the overall decision.

4 Baseline Acoustic Dialect Classification

GMM based acoustic classification is a popular method for text-independent dialect classification (Huang and Hansen, 2006) and therefore it is used as a baseline for our system. Fig. 2 shows the block diagram of the baseline gender-independent MFCC based GMM training system with 600 mixtures for each dialect. While testing, the incoming audio is classified as a particular dialect based on the maximum posterior probability measure over all the Gaussian Mixture Models. Mixture and frame selection based techniques as well as SVM-GMM hybrid techniques have been considered for dialect classification (Chitturi and Hansen, 2007). In order to assess the improvement by leveraging audio and text, we did not include these audio classification improvements in this study.

5 Dialect Classification using Language

As shown in Fig 1, the language based dialect classification module has two distinct classifiers. We describe in detail the n-gram and LSA based classifiers in the sections 5.1 and 5.2

5.1 N-gram based dialect classification

It is assumed that the text document is composed of many sentences. Each sentence can be regarded as a sequence of words **W**. The probability of generating **W** is given by $P(W|D) = P(w_1, w_2, \dots, w_m|D)$. Assuming the probability depends on the previous n words is $P(W|D) = \prod_{i=1}^{m} P(w_i|w_{i-n+1}, \dots, w_{i-1}, D)$ where m is the number of words in **W,** w_i is the word and D™ {UK, US, AU) is the dialect specific language model. The n-gram probabilities are calculated from occurrence counting. The final classification decision is given by C=$argmax_D \prod_{W\epsilon\varphi} P(W|D)$, where φ is a set of sentences in a document and D ™ {UK, US, AU}. In this study, we use the derivative measure of the cross entropy known as the test set perplexity for dialect classification. If the word sequence is sufficiently long, the cross entropy of the word sequence **W** is approximated as $H(W|D) = -\frac{1}{m} log_2 P(W|D)$. The perplexity of the test word sequence **W** as it relates to the language model D is $PP(W|D) = 2^{H(W|D)} = P(W|D)^{-1/m}$.The perplexity of the test word sequence is the generalization capability of the language model. The smaller the perplexity, the better

the language model generalizes to the test word sequence. The final classification decision is, C= $argmax_D \prod_{W \epsilon \varphi} P(W|D)^{-1/m}$, where is the set of sentences in a document, D ™ {UK, US, AU}.

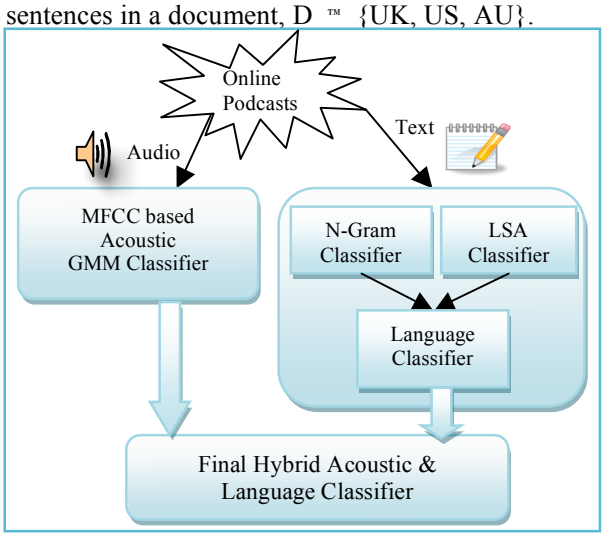

Figure 1: Proposed architecture

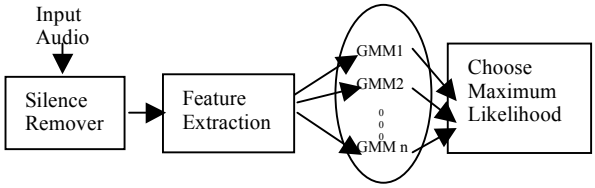

Figure 2: Baseline GMM based dialect classification

5.2 Latent Semantic Analysis for Dialect ID

One approach used to address topic classification problems has been latent semantic analysis (LSA), which was first explored for document indexing in (Deerwester et al., 1990). This addresses the issues of synonymy - many ways to refer to the same idea and polysemy – words having more than one distinct meaning. These two issues present problems for dialect classification as two conversations about a topic need not contain the same words and conversely two conversations about different topics may contain the same words but with different intended meanings. In order to find a different feature space which avoids these problems, singular value decomposition (SVD) is performed to derive orthogonal vector representations of the documents. SVD uses eigen-analysis to derive linearly independent directions of the original term by document matrix **A** whose columns correspond to the number of dialects, while the rows correspond to the words/terms in the entire text database. SVD decomposes this original term document matrix **A**, into three other matrices: $\mathbf{A=U*S*V^T}$, where the columns of **U** are the eigenvectors of $\mathbf{AA^T}$ (left eigenvectors), **S** is a diagonal matrix, whose diagonal elements are the singular values of **A,** and the columns of **V** are the eigenvectors of $\mathbf{A^TA}$(called right eigenvectors). The new dialect vector coordinates in this reduced 3 dimensional space are the rows of **V.** The coordinates of the test utterance is given by $\mathbf{q1=q^T*U*S^{-1}}$**.** The test utterance is then classified as a particular dialect based on the scores, given by the cosine similarity measure as $d_{best} = argmax_{d_i} \frac{(q_1.d_i)}{|q_1||d_i|}$, where d_i is one of the three dialects.

6 Results and Discussion

All evaluations presented in this section were conducted on the online podcast database described in the section 2. The first row of Table 3 shows the performance of the N-gram LM based dialect classification (69.1% avg. performance). From this we observe that this approach is good for US and UK, but not as effective for AU family dialect classification, with AU being confused with UK. The performance of the LSA based dialect classification is shown in the second row of Table 3. This classifier is consistent over all the dialects with better performance than the N-gram LM approach. There is more semantic similarity of US with AU than UK (24% vs 5% - false positives), while UK has a balanced semantic error with US and AU. This implies that there is more semantic information in these dialects than text sequence structure.

Next, the N-gram and the LSA classifiers are combined using optimal weights based on a greedy approach. Fig. 3 shows the performance of this hybrid classifier with respect to the weights of the individual classifiers (N-gram vs LSA: 0→all N-gram, 50→0.5 N-gram and 0.5 LSA, 100→ all LSA). After setting the optimal weights 0.18 to LSA and 0.82 to N-gram classifier, the hybrid classifier is seen to be consistent and better than the individual classifiers (Table 3: row 3 vs row2/row1). Performance of the hybrid classifier is not as good as the LSA classifier for AU classification, but significantly better for classification of US and UK. The hybrid classifier is better in all cases when compared to the N-gram classifier, with an overall average improvement of 7.3% absolute. The fourth row in Table 3 shows the performance of acoustic based dialect classification which is as good as the language based dialect classification, but it is noted that performance is poor for UK classification. It is expected that the type of errors made by text (word selection), semantics and acoustic space

will have differences and therefore we combine these acoustical and language classifiers as shown in Fig1. The overall performance of the proposed approach, combining the acoustic and language information, is better than the individual classifiers (Row 3 and Row 4 vs. Row 5 of Table 3). Even though the performance for US is reduced from 87.2% to 86.38%, the classification of UK is improved significantly from 54% to 74%. This shows that this approach is more consistent with accuracy that outperforms traditional acoustic classifiers with a relative improvement of 30%. With respect to a language only classifier, this hybrid classifier is better in all the cases.

7 Conclusions

In this study, we have developed a dialect classification (DC) algorithm that addresses family branch DC for English (US, UK, AU), by combining GMM based acoustic, and text based N-gram LM and LSA language information. In this paper, we employed LSA in combination with N-gram language models and GMM acoustic models to improve DC accuracy. The performance of the individual classifiers were shown to vary from 69.1%-72.4%. The final combined system achieves a DC accuracy of 79.5% and significantly outperformed the baseline acoustic classifier with a relative improvement of 30%, confirming that an integrated dialect classification system employing GMM based acoustic and N-gram LM, LSA based language information is effective for dialect classification.

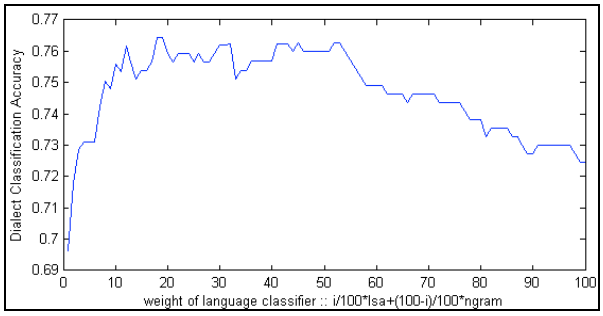

Figure 3: Language classifier

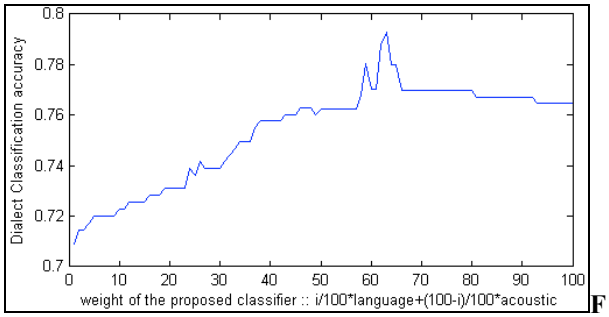

Figure 4: Acoustic + Language classifier

Accuracy→ Methods↓	US	UK	AU	Overall
N-Gram LM Classifier	75.2%	71.2%	60.7%	69.1%
Latent Semantic (LSA) Classifier	70.2%	68.5%	78.7%	72.47%
N-Gram+ LSA (Based on Text)	79.3%	74.6%	75.4%	76.4%
Acoustic GMM Classifier	87.2%	54.0%	73.3%	71.6%
Acoustic GMM + N-gram+ LSA	86.4%	74.6%	77.0%	79.5%

Table 3: Performance of classifiers on Dialect-ID

References

Diakoloukas, V.; Neumeyer, L.; Kaja, J.; 1997. "*Development of dialect-specific speech recognizers using adaptation methods"* IEEE- ICASSP

John Laver; 1994. "*Principles of Phonetics"*. Cambridge University Press, Cambridge, UK.

Gray, S.; Hansen, J.H.L; 2005. "*An integrated approach to the detection and classification of /dialects for a spoken document retrieval system"* IEEE- ASRU

Huang R; Hansen J.H.L.; 2005. *"Dialect/Accent Classification via Boosted Word Modeling,"* IEEE-ICASSP

Landauer,T.K., Foltz,P.W., & Laham,D; 1998. *" Introduction to Latent Semantic Analysis "* Discourse Processes, 25, 259-284.

Huang R; Hansen J.H.L.2007 *"Dialect Classification on Printed Text using Perplexity Measure and Conditional Random Fields,"* IEEE- ICASSP

Hasegawa-Johnson M, Levinson S.E, 2006 *"Extraction of pragmatic and semantic salience from spontaneous spoken English "* Speech Comm. Vol. 48(3-4)

Antoine, J.-Y 1996 *" Spontaneous speech and natural language processing. ALPES: a robust semantic-led parser "* ICSLP

Deerwester.S. et al.1990.*" Indexing by latent semantic anlysis "* Journal of American Society of Information Science, 391–407.

Chitturi. R, Hansen J.H.L, 2007.*" Multi stream based Dialect classification using SVM-GMM hybrids"* IEEE-ASRU

Huang. R, Hansen J.L.H.; 2006. *"Gaussian Mixture Selection and Data Selection for Unsupervised Spanish Dialect Classification"* ICSLP

Hansen J.H.L., Yapanel.U, Huang.R., Ikeno.A..; 2004. *"Dialect Anal'ysis and Modeling for Automatic Classification"* ICSLP

NIST- LRE 2005,*" Language Recognition Evaluation"*