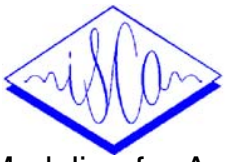

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Dialect Analysis and Modeling for Automatic Classification[†]

John H.L. Hansen, Umit Yapanel, Rongqing Huang, Ayako Ikeno

Robust Speech Processing Group, Center for Spoken Language Research
University of Colorado at Boulder; Boulder, Colorado 80309-0594, USA,
{jhlh, yapanel, huangr, ikeno@cslr.colorado.edu} web: http://cslr.colorado.edu/

Abstract

In this paper, we present our recent work in the analysis and modeling of speech under dialect. Dialect and accent significantly influence automatic speech recognition performance, and therefore it is critical to detect and classify non-native speech. In this study, we consider three areas that include: (i) prosodic structure (normalized f0, syllable rate, and sentence duration), (ii) phoneme acoustic space modeling and sub-word classification, and (iii) word-level based modeling using large vocabulary data. The corpora used in this study include: the NATO N-4 corpus (2 accents, 2 dialects of English), TIMIT (7 dialect regions), and American and British English versions of the WSJ corpus. These corpora were selected because the contained audio material from specific dialects/accents of English (N-4), were phonetically balanced and organized across U.S. (TIMIT), or contained significant amounts of read audio material from distinct dialects (WSJ). The results show that significant changes occur at the prosodic, phoneme space, and word levels for dialect analysis, and that effective dialect classification can be achieved using processing strategies from each domain.

1. Introduction

Every individual develops a characteristic speaking style at an early age that depends heavily on his language environment (i.e., the native language spoken), as well as the region where the language is spoken. Dialect classification is a recently emerging research topic in the speech recognition community since accent and dialect are some of the most important factors that influence performance in automatic speaker-independent speech recognition systems. *Dialect* is defined as "a regional variety of a language distinguished by pronunciation, grammar, or vocabulary." *Accent* on the other hand, is defined as "the relative prominence of a particular syllable or a word in pronunciation determined by a regional or social background of a speaker." The type of accent exhibited in second language pronunciation will depend on a number of speaker related factors such as; (i) the age at which a speaker learns the second language, (ii) the nationality of the speaker's language instructor, and (iii) the amount of interactive contact the speaker has with native talkers of the second language. It is known that dialect and foreign accent can have a significant impact on automatic speech recognition performance. For example, [1] showed that phonetic models built from foreign accented English are not less accurate than native ones at decoding data from the same accent. However, cross accented recognition experiments showed that phonemic models from a given accent were 1.8 times less accurate in recognizing speech from a different accent. Another study considered acoustic model development for decoding non-native speakers using MLLR to adapt prior models from 5 different languages[2]. Finally, another study considered the use of a polyphone decision tree to help improve speech recognition for non-native speakers[3]. These studies suggest that to improve speech recognition of non-native speech, it is necessary to develop effective automatic detection methods of dialect and accent.

A number of studies have shown that the acoustic space spanned by phonemes for American English will shift when a speaker is non-native, and that other factors such as voice-onset-time, voiced stop release time, duration, and pitch structure are also impacted[4,5]. Recent studies have also considered spectral (parametric and stochastic) trajectory modeling for accent classification in both closed and open accent scenarios[6,7]. Other studies have also considered the perception of accent through formal listener evaluations[8].

In this study, we focus on three domains which will contribute to analysis and modeling of effective techniques for dialect classification. The domains are (i) prosody structure using the N-4 corpus (because this corpus contains English speech from Canadian (English and French speakers), British, Dutch and German accented speakers), (ii) spectral acoustic space modeling and sub-word based classification using TIMIT (since this corpus contains 7 dialect regions across U.S.), and (iii) word based modeling/classification using N-4 and American and British English versions of the WSJ corpus.

2. Prosodic Based Analysis

Prosody has been shown to be a discriminative feature of languages and dialects in human perception [9], automatic language identification[10], and recognition and synthesis [11]. Prosodic characteristics that have been found to be discriminative include intonation and rhythmic structures of speech. This section considers English dialects and two prosodic features: normalized f0 range and syllable rate. The data set represents three types of native English dialects which are part of N-4 corpus[1]. Here, we focus on English produced by speakers from UK (United Kingdom), Canada (CA-1), and Canadian speakers whose native language is French (CA-2 {Montreal}). The speech samples were produced from a written passage, "the north wind and the sun." for a set of male speakers (e.g., speaker number: 7 UK, 10 CA-1, and 5 CA-2).

Normalized f0 Range: To obtain comparable measures of f0 range among speakers and dialect groups, the raw f0 values in Hz were converted into ratios. This was calculated for each sentence based on the average, maximum, and minimum f0 values of the utterance. For each speaker, the normalized f0 range values were averaged out based on all utterances. The final output is the average of all speakers in each dialect. Fig. 1 illustrates the results, where 0 is the normalized average f0, with positive/negative values being the max/min f0 range ratios. The result indicates that the f0 range ratio varies for each category of dialect, suggesting a potential dialect sensitive trait. To assess if this result was random, we analyzed further data from the N4 corpus using the same procedure. This data set contained speech samples from the following four categories: German, German-English, Dutch, and Dutch-English. Each category contained 10 male speakers' utterances. The results showed that the same speakers of German and German-English share the same f0 range

[†] This work was supported by U.S. Air Force Research Laboratory, Rome NY under Contract No. FA8750-04-1-0058.

10.21437/Interspeech.2004-38

properties. The maximum ratios were smaller and minimum ratios greater compared to the other languages or dialects in these data sets. The Dutch and Dutch-English speech also show the same trend. In other words, for the N-4 data the f0 range patterns are not random.

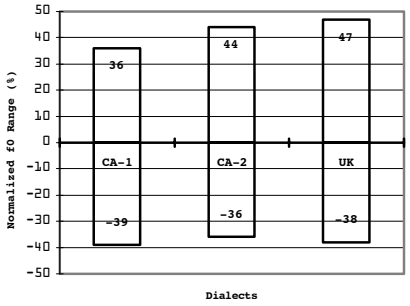

Fig. 1. Normalized f0 across Dialects of N-4 corpus.

Syllable Rate & Sentence Duration: The syllable rate (number of syllables per second) was estimated based on the average of the following three features: correlation between subband energy onsets, peaks in full-band energy, and spectral moment of the fullband energy trajectory. Fig. 2 shows average syllable rates and sentence durations of the first sentence per dialect group. Distinct dialect dependencies are clear. CA-1 speakers showed higher syllable rate with a corresponding shorter sentence duration. A more detailed analysis of the estimated syllable rate per speaker for all sentences confirms this observation (Fig. 3). The preliminary results suggest that prosodic characteristics carry discriminative features of dialect variations.

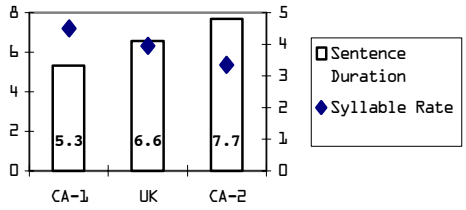

Fig.2: Sentence Duration (bars) and Syllable rate (diamonds)

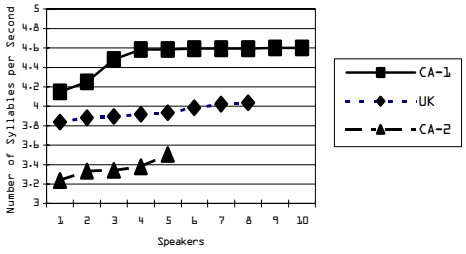

Fig.3: Syllable rate per speaker for 3 dialect/accents from N-4.

3. Subword Based Dialect Classification

Next, we consider spectral acoustic space experiments using subword-based classification strategies on the TIMIT corpus. The TIMIT database is divided into 8 *dialect regions*, (i.e., dr1...dr8), but dr8 is from speakers who moved frequently across the U.S., so it was excluded and our experiments used the first 7 American English (AE) dialects. We point out that since sentences from TIMIT were read speech, dialect traits may be different than that experienced from spontaneous speech. All experiments were conducted in a controlled fashion, were we trained dialect specific models and then scored each utterance with each dialect-specific model. Based-on the accumulated acoustic score obtained, we classified the dialect as the one that yields the maximum acoustic score.

3.1. Dialect-Adapted vs. Dialect-Trained Models

In order to perform dialect classification, we considered two methods: *dialect-adapted models* and *dialect-trained models*. The first approach is based on training a base model set and then adapting this base model to each dialect using available training data from that dialect. The second methods is to obtain dialect-trained models using only the available data for each dialect to train the models. For scoring, there are also several possible approaches. We can score each utterance separately, (i.e., *utterance-based scoring*) and then classify the dialect. Alternatively, we can use all available data from that speaker and decide the dialect for the input speaker, (i.e., *speaker-based scoring*). This second approach is clearly more desirable since the speaker is more likely to convey dialect dependent information over longer audio sequences. Using TIMIT dialects, we summarize our results for both training (Dialect Adapted & Dialect Trained) and scoring (Uttr-based & Spkr-based) strategies in Table 1. The best dialect classification rate was obtained with dialect-trained models and speaker-dependent scoring strategy (since there are 7 AE dialects, chance is 14%). Speaker-based scoring yields consistently better results than the utterance-based scoring regardless of the training strategy which is expected because: (i) there is more data in speaker-dependent scoring, and (ii) the dialect is tied to the speaker not to a single utterance. Pronunciation differences of certain phonemes we expect will be dialect dependent, but it is less likely that sufficient dialect discriminating phonemes will be consistently present for isolated word tests.

Training/Scoring[%]	Uttr-based	Spkr-based
Dlct-Adapted	24.39	28.66
Dlct-Trained	23.76	30.57

Table 1: *Dialect Classification Accuracy [%] for TIMIT task with two training and two scoring (utterance & speaker) methods.*

3.2. Inter-Dialect Confusion

Dialect classification based on read speech is a challenging task since word selection and sentence structure will not be a part of the speech. Therefore, basing the dialect decision only on the acoustic score is also difficult because there are a number of factors that impact this score such as speaker and background differences. In order to determine which dialects sound acoustically similar, we arranged our results with dialect-trained models and speaker-based scoring to construct an inter-dialect confusion matrix. This will allow us to evaluate the difficulty of the problem and also propose some solutions such as grouping close dialects together and potentially re-scoring in a *binary tree fashion*. Table 2 shows the confusion matrix, where the diagonal is the accuracy for each dialect. The **N.Total** row is the percentage of speakers, which are classified as a specific dialect. This analysis shows that approximately 60% (29.9+29.5) of the speakers are classified as either dr2 or dr4, whereas none of the speakers were classified as dr6. Thus, we can conclude that many of the TIMIT dialects are highly confusable to either dr2 or dr4.

3.3. Phone-based Scoring

We believe it is important to analyze the acoustic phone space for distinguishing dialects in speech recognition. The experiment from Sec. 3.2 was modified to determine the most discriminative

T/C	dr1	dr2	dr3	dr4	dr5	dr6	dr7
dr1	0.0	45.5	0.0	27.3	9.1	0.0	18.2
dr2	3.8	61.5	7.7	7.7	0.0	0.0	19.2
dr3	0.0	11.5	19.2	34.6	0.0	0.0	34.6
dr4	3.1	18.8	3.1	40.6	31.3	0.0	3.1
dr5	0.0	14.3	3.6	42.9	25.0	0.0	14.3
dr6	0.0	27.3	18.2	36.4	9.1	0.0	9.1
dr7	0.0	30.4	13.0	17.4	8.7	0.0	30.4
N.Total	0.9	29.9	9.3	29.5	11.9	0.0	18.4

Table 2: *Confusion Matrix for TIMIT Dialect classification rates [%] [T: True, C: Classified].*

phonemes in each dialect. Instead of using the total acoustic score, we use the acoustic score for each phone. This was done to identify which areas in the acoustic space phonemes show the smallest and largest intra-dialect variability. We performed this experiment and rank ordered the most discriminative phonemes and their classification accuracies for all 7 dialects. Table 3 shows *only* the first four most discriminative phonemes and the resulting dialect classification rates per AE dialect. Using the most discriminating phoneme for each dialect (1st column) we obtain a 12% absolute increase in the dialect classification rate (42.52% vs. 30.57 from Table 1). The results also show significant improvement for dr1 and dr6. Another encouraging result is that the discriminative phonemes are *not* overlapping, which means that, by using a weighting scheme, we may combine the acoustic score from each phone appropriately in order to maximize the classification rate.

Ph/Acc	1st	2nd	3rd	4th
dr1	IH/36.4	OY/36.4	Y/36.4	BD/27.3
dr2	EY/34.6	DD/26.9	D/26.9	DX/26.9
dr3	M/34.6	Z/34.6	IX/30.8	KD/30.8
dr4	L/40.6	R/37.5	SH/37.5	ER/28.1
dr5	AA/32.1	UW/28.6	DH/25.0	ER/25.0
dr6	IX/45.5	Y/27.3	AA/18.2	AE/18.2
dr7	BD/73.9	ZH/69.6	PD/65.2	TH/47.8
Overall	42.52	37.27	34.28	29.15

Table 3: The first 4 most discriminative phonemes and their accuracy [%] for the classification of different dialects.

4. WDC: Word based Dialect Classification

In this third area, we note that human listeners can make reasonable decisions on English dialects based on isolated words. Individual words do encode high-level features such as formant and intonation so that it can be useful for dialect classification. Sentence level dialect classification can also be fulfilled at the word level. If the highest percent words are classified correctly (in two dialects classification, it is 50%), the sentence will be classified correctly. Although the vocabulary size of English is extremely large, almost each sentence includes some of the most frequently occurring words. Greenberg reported that one hundred words account for 66% of the individual tokens in the SwitchBoard corpus, which has 26k distinct words [12]. If these words are dialect dependent, we would only need to train models for the most frequently occurring words and this should provide useful dialect information. Since only a few words within a sentence are enough to make a decision, the requirement on the speech recognizer is very low. A simple Wall Street Journal (WSJ) AE acoustic models and language models are used for all English dialects experiments in this section. Although the WER is high, it can still supply sufficient information to the dialect classifier.

Fig. 4: *Block Diagram of WDC Training*

The WDC training stage (Fig. 4) requires audio, word level transcripts and gender information if a gender-dependent scheme is preferred. Viterbi forced alignment is used to obtain word boundaries automatically. Each word is extracted from audio and all segments belonging to that word are grouped together. Female, male and gender-independent GMMs are trained for each word using 39-dimensional MFCC vectors.

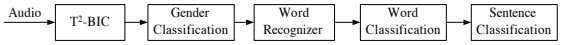

Fig. 5: *Block Diagram of WDC Testing*

Fig. 5 shows the WDC test stage. In the unsupervised context, a modified T^2-BIC [13,15] scheme is used to segment the audio based on speaker turns. In the supervised context, the segments are known. Each segment is submitted into a GMM gender classifier to decide the speaker gender of that segment. The Sonic [13] speech recognizer with WSJ acoustic models and language models are used to decode the word sequence of the segments. Since recognition requirements are very low, this simple recognizer is used for all English dialects and corpora without any adaptation. The recognizer generates the hypothesis text (word) sequence with corresponding confidence scores. A GMM classifier is applied to classify the words that have high confidence scores and corresponding models. If no gender dependent model exists, a gender-independent model is employed. Sentence level dialect decision is made based on the dialect class with the most associated words.

Fig. 6 shows probe word classification results with the common English paragraph spoken in Canadian-French accent, German accent, Dutch accent from the NATO N-4 corpus [3]. The x-axis is the number of phonemes in the word; the y-axis is the word classification rate. We see that all words can be classified with high accuracy, above 70%, except single-phoneme words. Since the word classification rate is more than 50%, Fig. 6 suggests that two-or-more-phoneme words are efficient for accent/dialect classification at the sentence level.

4.1: Dialect Experiments: WSJ, N-4

Having established that word level dialect classification can be effective, we consider a more extensive set of experiments using WSJ, WSJCAM0 and NATO N4 corpora for training and testing. The WSJ and WSJCAM0 are the Wall Street Journal read by American English speakers and British English speakers respectively. The NATO N-4 corpus includes 4 kinds of dialects or accents of American English produced by British, Canadian, German and Dutch speakers. The test data is bout 100 minutes in duration. All other data is used for training. A pair of Broadcast news GMMs trained using other data is applied for gender classification. The confidence level of 0.95 is used as the threshold for selecting words from the automatic generated transcripts. Since the GMM is used for word modeling, our baseline system is a phoneme level GMM dialect classifier (PDC: Phoneme based Dialect Classification).

Table 4 shows how the WDC algorithm compares with the PDC baseline in a supervised (top 2 rows) and unsupervised (bottom 2 rows) context. The two dialects are: American and British dialects of English in WSJ-WSJCAM0 corpora, and the four dialects/accents: British, Canadian, German, Dutch accent English in the N4 corpus. In the WSJ-WSJCAM0 test, the phoneme level GMMs (PDC) cannot classify the data at all: the *posteriori* probabilities of the British English GMMs are higher than those of American English GMMs. This implies that

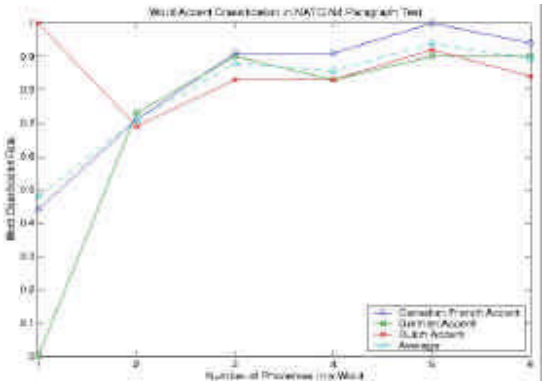

Fig. 6: *Word Accent Classification Rate in N4 Paragraph Test*

classification of British and American English is difficult only with phoneme level information. From Table 4 shows that the algorithm based on word modeling (WDC) is more successful, with the 4-accent N4 corpus achieving a 100% dialect classification accuracy. Unsupervised classification still maintains high classification accuracy, noting that T^2-BIC achieves 3.5% miss rate with 9.1% false alarm in speaker turn detection with the WSJ-WSJCAM0 test data (i.e., T^2-BIC is an effective pre-processing algorithm for unsupervised dialect classification).

Table 5 shows WER for the word recognizer in the WDC, using WSJ American English acoustic and language models. The recognizer employs a simple pass without any adaptation. Therefore, the WER for "American-WSJ" is very low and high for others (especially N4 data). WDC can still achieve high dialect classification even with a WER as high as 84.6%. This observation proves our argument above: that since a small number of words within a sentence are enough for making a decision, the requirement on the recognizer is very low. Since "American-WSJ" achieves a much lower WER than "British-WSJ", the "American-WSJ" will have a higher recall rate than the "British-WSJ" which is confirmed by Table 4. The WER of the 4 dialects/accents from N4 is almost the same, so the WER does not impact the dialect classification rate. That may be the reason why the two-accent classification accuracy is lower than the four-dialect classification accuracy.

Scheme-Supervised	**PDC**	**WDC**
WSJ-WSJCAM0	-	90.5%
NATO N4	66.7%	100%
Scheme-Unsupervised	**PDC**	**WDC**
WSJ-WSJCAM0	-	85.9%
NATO N4	59.7%	93.4%

Table 4: *Word Dialect Classification vs. Phoneme Dialect Classification for Supervised (top 2 rows), and Unsupervised (bottom 2 rows) methods.*

English Dialect /Accent	*American-WSJ*	*British-WSJ*	*British-N4*
WER	3.8%	26.5%	101.1%
Canadian-N4	*German-N4*	*Dutch-N4*	***N4-Total***
74.2%	88.7%	77.8%	84.6%

Table 5: *WER of Word Recognizer in WDC*

Accent	Precision	Recall
American-WSJ	81.68%	99.68%
British-WSJ	99.72%	84.01%

Table 6: *Precision and Recall of the WSJ-WSJCAM0 data with WDC (in supervised context)*

5. Conclusions

In this study, we have considered three areas for dialect modeling and classification. First, prosodic analysis showed that normalized f0 range, syllable rate, and sentence duration are promising characteristics for dialect modeling. Second, we considered acoustic phone space analysis to determine which phones were most discriminating relayer of dialect. Using entire phoneme space based models, we saw that some American English dialects were not easily identified, however by selecting anyone of the top 4 discriminating phonemes increased classification rates by 12% absolute. Next, we considered a word based Dialect Classification (WDC) scheme which was shown to provide effective performance in both supervised and unsupervised English dialect classification scenarios. The word recognition requirements within the WDC is very low, so effective dialect classification can be performed using a common ASR engine to make the dialect decision. The WDC can be incorporated into other speech tasks such as LID. In the future, we will employ HMMs to the word modeling instead of GMMs in order to model the transition of the phoneme sequence of the word. Also, the word classification contribution will be weighted based on the dialect discriminative power of the words. Finally, we will consider a parallel set of listener evaluations to benchmark against automatic dialect classification.

REFERENCES

[1] A.D. Lawson, D.M. Harris, J.J. Grieco, "Effect of Foreign Accent on Speech Recognition in the NATO N-4 Corpus," Eurospeech-2003, pg. 1505- 1508, Geneva, Swiss., Sept. 2003.

[2] V. Fischer, E. Janke, S. Kunzmann, "Recent Progress in the Decoding of Non-Native Speech with Multilingual Acoustic Models," Eurospeech-03, pg. 3105-3108, Swiss., Sept. 2003.

[3] Z. Wang, T. Schultz, "Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization," Eurospeech-2003, pg. 1449- 1452, Geneva, Swiss., Sept. 2003.

[4] L.M. Arslan, J.H.L. Hansen, "A study of Temporal Features and Frequency Characteristics in American English Foreign Accent," J. Acoustical Soc. of America, **102**(1), pp. 28-40, July, 1997.

[5] L.M. Arslan, J.H.L. Hansen, "Language Accent Classification in American English," Speech Communication, vol. 18(4), pp. 353-367, July 1996.

[6] P. Angkititrakul, J.H. L. Hansen, "Use of Trajectory Models for Automatic Accent Classification," Eurospeech-03, pp.1353-1356, Geneva, Switzerland, Sept. 2003.

[7] P. Angkititrakul, J.H.L. Hansen, "Stochastic Trajectory Model Analysis For Accent Classification," ICSLP-2002: Inter. Conf. Spoken Lang. Proc., vol. 1, pp. 493-496, Denver, CO, Sept. 2002.

[8] Flege, J.E., Munro, M.J., MacKay, I.R.A., "Factors affecting strength of perceived foreign accent in a second language," *J. Acoustical Society America*, **97**(5):3125 –3134, 1995.

[9] M. Barkat, J. Ohala, F. Pellegrino, "Prosody as a distinctive feature for the discrimination of Arabic dialects", Eurospeech-99, p.395-398, 1999.

[10] J.L. Rouan, J. Farinas, F. Pellegrino, "Automatic modeling of rhythm and intonation for language identification", Inter. Congress Phonetic Sciences (15th ICPhS), pp. 567-570, 2003.

[11] Q. Yan, S. Vaseghi, "A comparative analysis of UK and US English accents in recognition and synthesis", IEEE ICAASP-90, vol.1, pp. 413-416, 2002.

[12] S. Greenberg, "On the Origins of Speech Intelligibility in the Real World", *ESCA Workshop Robust Speech Recognition for Unknown Communication Channels*, pp. 23-32, France, 1997.

[13] R. Huang, J.H.L. Hansen, "Advances in Unsupervised Audio Segmentation for the Broadcast News and NGSW Corpora", *ICASSP*, Montreal, Canada, May, 2004.

[14] B. Pellom, "Sonic: the University of Colorado Continuous Speech Recognizer", TR-CSLR-2001-01, Univ. Colorado, March, 2001

[15] B. Zhou, J.H.L. Hansen, "Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion", *ICSLP*, (1):714-717, Beijing, China, 2000.