DIALECT/ACCENT CLASSIFICATION VIA BOOSTED WORD MODELING

Rongqing Huang, John H.L. Hansen

Robust Speech Processing Group, Center for Spoken Language Research
University of Colorado, Boulder, CO, USA
{huangr,jhlh}@cslr.colorado.edu

ABSTRACT

This paper addresses novel advances in English dialect/accent classification/identification. A word level based modeling technique is proposed that is shown to outperform a LVCSR based system with significantly less computational costs. The new algorithm, which is named WDC (Word based Dialect Classification), converts the text independent decision problem into text dependent problem and produces multiple combination decisions at the word level rather than make a single decision at the utterance level. There are two sets of classifiers employed for WDC: word classifier $D_{W(k)}$ and utterance classifier D_u. $D_{W(k)}$ is boosted via real AdaBoost.MH algorithm in the probability space directly instead of the feature space. D_u is boosted via the dialect dependency information of the words. Two dialect corpora are used in the evaluation. Significant improvement in dialect classification is achieved for both corpora.

1. INTRODUCTION

In order to achieve a reasonable identification accuracy in English dialect/accent classification, it is first necessary to understand how dialects differ. Fortunately, there are plenty of studies on English dialectology [12, 16, 17]. The English dialects differ in the following ways[17]:

1. Phonetic realization of vowels and consonants
2. Phonotactic distribution (e.g., rhotic in *farm*: /farm/ vs. /fa:m/)
3. Phonemic System (the number or identity of phonemes used)
4. Lexical Distribution
5. Rhymical characteristics:
   - syllable boundary (e.g., *self#ish* vs. *sel#fish*)
   - pace (average number of syllables uttered per second)
   - lexical stress
   - intonation (sentence level, semantic focus)
   - voice quality (e.g., creaky voice vs. breathy voice)

The first 4 points are visuable at the word level. All the rhymical characteristics except intonation can be, or at least partially, represented at the word level [17]. In [12], a single word "hello" was used to distinguish 3 dialects in American English. From the linguistic point of view, a word may be the best unit to classify dialects. For an automatic classification system, it is impossible to build models for all possible words from different dialects. Fortunately, the words in a language are very unevenly distributed. The 100 most common words account for 40% of the occurrences in the Wall Street Journal (WSJ) corpus, which has 20K distinct words [8], and account for 66% in the SwitchBoard corpus, which has 26K distinct words [4]. So only a small set of words require modeling. In [6, 8, 13], word level information was embeded into the phoneme models and improvement in language identification was achieved. In this paper, a system based solely on word models is implemented and shows to have great advantage over a Large Vocabulary Continuous Speech Recognition (LVCSR)-based system,which is claimed to be the best performing system in Language Identification [18].

AdaBoost algorithm [3] is a powerful learning algorithm. In [1, 2, 10], researchers applied the AdaBoost algorithm into GMM/HMM modeling and obtained consistent but small improvement with large computational costs. In this paper, several AdaBoost variations are compared and the best one is applied to the word classifier $D_{W(k)}$ directly instead of model re-training. This method obtains significant improvement with small computational cost. The dialect dependency of words is also considered and embeded into the WDC through the utterance classifier D_u.

2. BASELINE CLASSIFICATION SYSTEM

It is known that LVCSR-based systems achieve high performance in language identification since they use knowledge

This work was supported by US Air Force

from phoneme and phoneme sequence to word and word sequence [18]. In [5, 9, 15], the LVCSR-based systems were shown to perform well in language identification. We implement a similar LVCSR-based system as our dialect classification baseline system. Fig. 1 shows a block digram of the system, where N is the number of dialects. AM_i, LM_i are the acoustic model (trained on triphones) and the language model (trained on word sequences) of dialect i respectively. L_i is the likelihood of dialect i. The final decision is obtained as:

$$D_L = \arg\max_i \ L_i, \ i = 1, 2, \ldots, N. \tag{1}$$

The LVCSR-based system requires significant word level transcripted audio data to train the acoustic and language models for each dialect. During the test phase, N recognizers are employed in parallel. It is among the most computationally complex algorithms and achieves very high dialect classification accuracy.

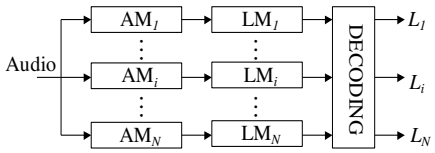

Fig. 1. *LVCSR-based Dialect Classification System*

3. WDC AND EXTENSIONS

3.1. Basic WDC Algorithm

Fig. 2 is the block diagram of WDC system training. For dialect i, the audio data A_i and the word level transcript T_i are given. Viterbi forced alignment is applied to obtain the word boundaries. The data that comes from the same word is grouped together. Common words across all the dialects are kept and an HMM is trained for each word and each dialect. The set of common words is $\mathcal{J}$. So the set of HMMs is summarized as,

$$\Psi = \{HMM_{ij}\}, \ i = 1, 2, \ldots, N, \ j \in \mathcal{J}, \tag{2}$$

where N is the number of dialects. The transcript $\mathcal{T} = \{T_1, \ldots, T_i, \ldots, T_N\}$ is used to train a language model $\overline{LM}$, which is used in the recognizer during the classification.

Fig. 3 is the block diagram of the WDC test. A gender classifier is applied to the input utterance if the gender-dependent classification is preferred. The gender classifier is a GMM classifier trained with Broadcast News data, which is used in our other studies. Usually, the dialect data is not large enough to train a robust acoustic model. Also, acoustic modeling is very time consuming. So a previously well-trained decision tree triphone model AM_p is

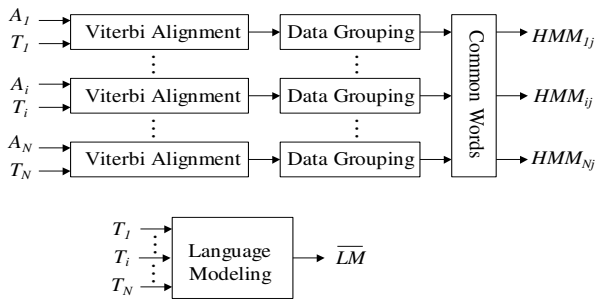

Fig. 2. *Block Diagram of WDC Training*

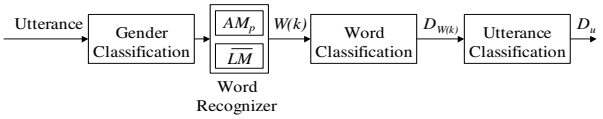

Fig. 3. *Block Diagram of WDC Test*

used, which is independent of the dialect data. The data-specific language model $\overline{LM}$ is intentionally used to force the word recognizer to output the words which have models trained before. The word recognizer therefore outputs the word sequence $\mathbf{O}$ with boundary information. The effective word sequence $\mathbf{W}$ is

$$W(i) \leftarrow O(i), \ if \ O(i) \in \mathcal{J}, \ i = 1, 2, \ldots, \tag{3}$$

The word classification is based on a Bayesian classifier, where the decision $D_{W(k)}$ is

$$\begin{aligned} D_{W(k)} &= \arg\max_i Pr(W(k)|HMM_{iW(k)}), \\ W(k) &\in \mathcal{J}, \ k = 1, 2, \ldots, K, \ i = 1, 2, \ldots, N, \end{aligned} \tag{4}$$

where N is the number of dialects, $\mathcal{J}$ is the set of common words across N dialects, $Pr(\cdot|\cdot)$ is the conditional probability, $K = |\mathbf{W}|$ is the size of the effective word sequence $\mathbf{W}$. The final decision on the utterance is,

$$D_u = \arg\max_i \sum_{k=1}^{K} \mathcal{I}(D_{W(k)}, i), \ i = 1, 2, \ldots, N. \tag{5}$$

Here, $\mathcal{I}(\cdot, \cdot)$ is the indicator function defined as,

$$\mathcal{I}(f, g) = \begin{cases} 1 & f = g \\ 0 & f \neq g \end{cases} . \tag{6}$$

If we compare Eq. 1 with Eq. 4 and 5, we see that WDC turns the single text-independent decision problem at the utterance level into a multiple combination text-dependent decision problem at the word level. WDC also provides options for further modeling and decision space improvement.

3.2. Boosting Classifier $D_{W(k)}$ in the Probability Space

Let us consider Eq. 4 first. For simplicity, let us represent the word sequence $W(k)$ and HMM as,

$$m \leftarrow W(k),\ \Theta \leftarrow HMM, \tag{7}$$

and define a probability vector,

$$\mathbf{p}^m = [Pr(m|\Theta_{1m})\ Pr(m|\Theta_{2m})\ \ldots Pr(m|\Theta_{Nm})], \tag{8}$$

with a hypothesis function as follows,

$$h(\mathbf{x}) = \arg \max_{1 \le i \le |\mathbf{x}|} x_i. \tag{9}$$

With this, we can represent Eq. 4 as,

$$D_m = h(\mathbf{p}^m), \tag{10}$$

without loss of generality, the word label term m is dropped, so as to obtain the following relations,

$$D = h(\mathbf{p}). \tag{11}$$

Apparantly, this decision strategy does not apply the classification information of the training samples. Given the training samples $(\mathbf{p}_j, y_j)$, where $y_j = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, T$, T is the total number of training samples of the word m in the N dialects, the AdaBoost algorithm can be applied to learn a sequence of "base" hypotheses h_t and the "vote power" α_t of each hypothesis in the final classifier. The boosted $D_{W(k)}$ is

$$D_{W(k)} = D = \sum_{t=1}^{n} \alpha_t h_t(\mathbf{p}). \tag{12}$$

The implementation details of AdaBoost are discussed in [3, 14]. Here, n is the number of iterations and usually goes to several hundred for convergence. This reflects another motivation for us to boost the classifier in the probability space instead of the feature space as in [1, 2, 10]. The latter results in HMM training for each iteration, and is computationally expensive.

3.3. Boosting Classifier D_u via Dialect Dependency

The words usually encode different levels of dialect dependency information. That is, the words do not have the same "decision power" in Eq. 5. A new boosted version of our classifier D_u can be formed as follows,

$$D_u = \arg\max_i \sum_{k=1}^{K} \mathcal{I}(D_{W(k)}, i) \cdot l_{W(k)\cdot i},\ i = 1, 2, \ldots, N, \tag{13}$$

where $l_{W(k)\cdot i}$ is the measure of dialect dependency which is defined as,

$$l_i = \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \left\{ \frac{1}{T_i} \sum_{t=1}^{T_i} \log \frac{Pr(X_{it}|HMM_i)}{Pr(X_{it}|HMM_j)} + \frac{1}{T_j} \sum_{t=1}^{T_j} \log \frac{Pr(X_{jt}|HMM_j)}{Pr(X_{jt}|HMM_i)} \right\}. \tag{14}$$

For simplicity, the word label term $W(k)$ is dropped here, where T_i is the number of training samples of word $W(k)$ in dialect i; X_{it} is the t^{th} training sample in dialect i, $i = 1, 2, \ldots, N$. l_i can be computed during the training stage, so there is no additional computational cost for evaluation.

4. EXPERIMENTS

The speech recognizer used in our studies is Sonic system [11], which trains decision-tree triphone acoustic model and back-off trigram language model. The feature used in the study is MFCC (static, delta, double delta).

4.1. Evaluation Corpora

Two corpora are used for evaluation. WSJ American and British English corpus (WSJ and WSJCAM0), and NATO N4 Foreign Accent of English corpus [7]. Table 1 shows the information of training and test sets for the corpora.

4.2. Performance of Boosted Classifier $D_{W(k)}$

In order to determine the proper number of iterations for AdaBoost, the original training set is partitioned into 3 : 1 training and test sets. In order to obtain robust classifiers, only the word which has enough training samples is boosted. Table 2 shows the information of boosted models. The occurrence coverage is computed in the training set. Fig. 4 shows the error rate of $D_{W(k)}$ in the newly partitioned WSJ training and test sets. From Fig. 4, two observations can be made: first, the real AdaBoost.MH is the best AdaBoost algorithm, which is consistent with the original paper [14]; second, hundreds of iterations are necessary for convergence.

Table 2. *AdaBoost applied on the Corpora*

Data	Word Models	Boosted Models	Model Coverage	Occurrence Coverage
WSJ	1642	51	3%	44%
N4	129	7	5%	22%

4.3. Evaluation on the WDC and Extensions

Sec. 4.2 shows that the AdaBoost algorithm can boost $D_{W(k)}$ significantly. Now the boosted classifiers are applied to

Table 1. *The Two Evaluation Corpora*

Data	Training Set				Test Set			Dialects/
	Vocabulary	Speakers	Size	style	Speakers	Size	Style	Accents
WSJ	20K	375	40 hours	read	22	1 hour	read	2(American,British)
N4	1159	211	22 hours	read/Spon.	31	43 mins	read/spon.	4(British,Canadian,Dutch,German)

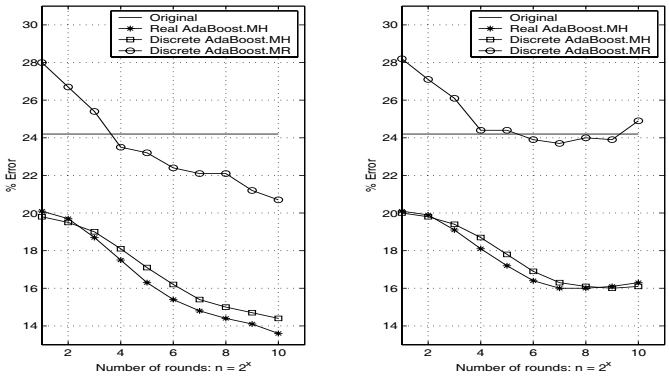

Fig. 4. *Training (left) and Test (right) Error using 3 AdaBoost Methods on WSJ Gender-Independent HMMs. Dialect classification error versus number of AdaBoost iterations($n = 2^x$)*

the basic WDC (WDC+AB, Sec. 3.2). The dialect dependency can also "boost" D_u classifier (WDC+DD, Sec 3.3). $D_{W(k)}$ and D_u can be boosted simultaneously and obtain WDC+AB+DD. There are no specific parameters required in the WDC algorithm and its extensions. The baseline system is LVCSR-based. The length of the test utterance is 9 seconds.

Table 3. *Classification Error(%) of Algorithms*

Data	LVCSR	WDC	WDC+ DD	WDC+ AB	WDC+ AB+DD
WSJ	5.9	3.1	2.7	2.1	1.9
N4	5.5	3.4	1.9	3.0	1.6

From Table 3, the basic WDC significantly outperforms the LVCSR-based system, which has been claimed to be the best performing system in language identification. The WDC requires much less computation, especially in the evaluation stage since only one recognizer is used instead of N parallel recognizers. The word classifier $D_{W(k)}$ can be directly boosted by the AdaBoost algorithm in the probability space, and the utterance classifier D_u can be boosted by dialect dependency. These extensions of the basic WDC also show great performance. Since only a few word models in the N4 corpus are boosted (see Table 2), the "WDC+AB" in N4 corpus does not show the same level of improvement as in WSJ corpus.

5. CONCLUSIONS AND FUTURE WORK

An effective word based dialect classification technique called WDC is proposed. A direct comparison between a LVCSR-based dialect classifier versus WDC shows that WDC achieves better performance with less computational and memory requirements. The basic WDC algorithm also offers a number of areas for extensions. The AdaBoost algorithm and dialect dependency are embeded into the word classifier $D_{W(k)}$ and utterrence classifier D_u respectively. Further improvement is achieved with these extensions. In the future, we plan to test the boosted $D_{W(k)}$ and D_u on large multiclass corpora.

6. REFERENCES

[1] C.Dimitrakakis, S.Bengio, "Boosting HMMs with an Application to Speech Recognition", *ICASSP*, vol.5, pp.621-624, Montreal, Canada, May, 2004

[2] S.W.Foo, L.Dong, "A Boosted Multi-HMM Classifi er for Recognition of Visual Speech Elements", *ICASSP*,vol.2, pp.285-288, Hong Kong, China, Apr., 2003

[3] Y.Freund, R.E.Schapire, "A Decision-Theoretic Generalization of on-line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, 55(1):119-139, 1997

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

[5] J.L.Hieronymus, S.Kadambe, "Robust Spoken Language Identifi cation using Large Vocabulary Speech Recognition", *ICASSP*, vol.2, pp.1111-1114, Munich,Germany, Apr., 1997

[6] S.Kadambe, J.L.Hieronymus, "Language Identifi cation with Phonological and Lexical Models", *ICASSP*, vol.5, pp.3507-3510, Detroit, USA, May, 1995

[7] A.Lawson, D.Harris and J.Grieco, "Effect of Foreign Accent on Speech Recognition in the NATO N-4 Corpus", *EuroSpeech*, pp. 1505-1508, Geneva, Switzerland, Sept., 2003

[8] D.Matrouf, M.Adda-Decker, L.F.Lamel, J.L.Gauvain, "Language Identifi cation Incorporating Lexical Information", *ICSLP*, vol.2, pp.181-185, Sydney, Australia, Dec., 1998

[9] S.Mendoma, L.Gillick, Y.Ito, S.Lowe, M.Newman, "Automatic Language Identifi cation using Large Vocabulary Continuous speech Recognition", *ICASSP*, vol.2, pp.785-788, Atlanta, USA, May, 1996

[10] C.Meyer, "Utterance-Level Boosting of HMM Speech Recognizers", *ICASSP*, vol.1, pp.109-112, Orlando, USA, May, 2002

[11] B.Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer", *Tech. Report TR-CSLR-2001-01*, University of Colorado, USA, March 2001.

[12] T.Purnell, W.Idsardi, J.Baugh, "Perceptual and Phonetic Experiments on American English Dialect Identifi cation", *Journal of Language and Social Psychology*, vol.18, no.1, pp.10-30, March, 1999

[13] P.Ramesh, E.Roe, "Language Identifi cation with Embedded Word Models", *ICSLP*, vol.4, pp.1887-1890, Yokohama, Japan, Sep., 1994

[14] R.E.Schapire, Y.Singer, "Improved Boosting Algorithms using Confi dence-Rated Predictions", *Machine Learning*, vol.37, no.3, pp.297-336, 1999

[15] T.Schultz, I.Rogina, A.Waibel, "LVCSR-Based Language Identifi cation", *ICASSP*, vol.2, pp.781-784, Atlanta, USA, May, 1996

[16] P.Trudgill, "The Dialects of England", 2nd edition, *Blackwell Publishers Ltd*, Oxford, UK, 1999

[17] J.C.Wells, "Accents of English", vol. I, II, III, *Cambridge University Press*, Cambridge, UK, 1982

[18] M.A.Zissman, K.M.Berkling, "Automatic Language Identifi cation", *Speech Communication*, vol.35, pp.115-124, 2001