HYPOTHESIS-DRIVEN ACCENT IDENTIFICATION

Laura Mayfield Tomokiyo

Carnegie Mellon University

ABSTRACT

Native and non-native use of language differs, depending on the proficiency of the speaker, in clear and quantifiable ways. It has been shown that customizing the acoustic and language models of a natural language understanding system can significantly improve handling of non-native input; in order to make such a switch, however, the nativeness status of the user must be known. In this paper, we show how the recognition hypothesis can be used to predict with very high accuracy whether the speaker is native. Effectiveness of both word-based and phone-based classification are evaluated, and a discussion of the primary discriminative features is presented. In an LVCSR system in which users are both native and non-native, we have achieved a 15% relative decrease in word error rate by integrating this classification method with speech recognition.

1. INTRODUCTION

Recognition performance on non-native speech can be significantly poorer than on condition-matched native speech. A variety of methods have been proposed for adapting acoustic and lexical models to non-native speech; most approaches, however, assume prior knowledge that the speaker is non-native. An algorithm for detecting non-native speech is therefore needed if non-native modeling is to be fully integrated into an LVCSR system.

In this paper, we present a hypothesis-driven method for identifying non-native speakers. Using recognizer output of either words or phonemes, we apply Bayesian classification to determine whether or not the speaker is native. This approach has the advantage of being independent of the recognizer-internal representation of acoustic features, and it requires no additional training of acoustic or language models.

Bayesian classification is well suited to this task for several reasons. Bayesian learning methods support probabilistic hypotheses, which allow would a nativeness threshold to be set or the result to be incorporated with other sources of information. Bayesian classification incorporates the marginal probability of the class, so knowledge of the distribution of speakers likely to use the system can help to improve classification accuracy. Bayesian models also handle conflicting examples gracefully, and are not as vulnerable to data sparsity problems as partitioning methods like decision tree learning.

Work in accent discrimination has focused primarily on acoustic features. Fung and Liu [1] have reported success in discriminating native- and Cantonese-accented English using energy and formant observations in a hidden Markov model (HMM). Teixeira, Trancoso, and Serralheiro [2] also used HMMs, but in a configuration more often associated with language identification, in a six-way accent identification task. In their framework, individual phoneme model sets were trained for each accent-language pair and integrated as sub-nets of a larger HMM. Methods that use acoustic features, however, can be difficult to implement if the features are not in a form that is readily accessible to the researcher. Approaches involving training of acoustic models are also time consuming and computationally expensive.

The method we describe is extremely fast and requires neither linguistic knowledge nor feature extraction. Hypotheses from a general English recognizer are classified as native or non-native by a naive Bayes classifier that has been trained on examples of native and non-native speech. If the speaker is found to be non-native, the utterance is re-evaluated using a customized acoustic model for optimal recognition accuracy.

2. BAYESIAN CLASSIFICATION

2.1. Overview

A Naive Bayes classifier incorporates information about statistical priors on the target classes as well as the features present in each example. A test example is classified by assigning it to the class calculated as most likely to have produced it. For an utterance u which may assigned to a class c, its score is calculated as follows:

$$P(c|u) = \frac{P(c)P(u|c)}{P(u)}$$

When choosing between classes, we need not calculate the probability directly, since we want only to find the maximum score, and $P(u)$ is constant across classes.

$$argmax_{c_i} P(c_i|u) = argmax_{c_i} P(c_i)P(u|c_i)$$

We cannot gather enough training data to reliably estimate the probability of every possible utterance for arbitrary native and non-native speakers. Even simplifying to seqeunces of POS-tags will not permit completely reliable estimation of probabilities of utterances. Instead, we take advantage of the fact that we can break utterances down into their constituent n-grams, and make the simplifying assumption that constituent n-grams are statistically independent, and that their order is unimportant. These assumptions leave us with the following formula in the trigram case:

$$P(u|c_i) \approx \prod_{w_a w_b w_c \epsilon u} P(w_a w_b w_c|c_i) \quad (1)$$

leading to the selection of class c_i according to

$$c_i = argmax_{c_i} P(c_i) \prod_{w_a w_b w_c \epsilon u} P(w_a w_b w_c|c_i) \quad (2)$$

In our framework, the class c_i is the nativeness status of the user. The tokens w_i are words, phonemes, parts of speech, or phone classes, as we will see in Sections 4 and 5. Based on the token sequences that are seen in the training hypotheses, a naive Bayes classifier is trained; based on the token sequences seen in the run-time hypotheses, the classifier determines whether the speaker is native and non-native, finding the most probable class c_i given the hypotheses.

2.2. Implementation and execution

The Rainbow statistical text classification package [3] was used for all classification experiments. Rainbow implements a naive Bayes classifier for text, with a number of features specialized for text applications.

To frame accent identification as a text classification problem, each set of utterances from one training speaker was treated as a document. From one set of documents labeled as native and another set labeled as non-native, the classifier learns features distinguishing them and is able to predict whether a new document is native or non-native. This formulation of the task prevents the model being trained on the idiosyncrasies of any one speaker, and allows very straightforward execution.

In building and testing the model, no feature selection, or vocabulary specialization, was used. Stop-words were *not* excluded, as they were found to be effective discriminators. The model takes less than one second to build in all of the configurations we describe. Data was randomly partitioned into 70% training and 30% testing, with the results averaged over 20 runs.

3. SPEECH RECOGNITION

3.1. Speech data

This paper reports on classification and recognition results from an English read news task, using ten native speakers of Japanese and eight native speakers of American English. Each speaker read three articles, one of which was read by all speakers and two of which were read only by that speaker. Each article contained approximately 50 sentences. The ten non-native speakers were all of similar English proficiency, and had had similar degrees of exposure to English. This data set is described in more detail in [4].

3.2. Recognizer

The JRTk speech toolkit [5] was used to produce recognizer hypotheses and to evaluate overall system performance. WER of this system on Broadcast News F0 data is 9.4%. On the speakers in these experiments, WER was 21% for native speakers and 58% for non-native speakers. It has been established that the lower performance for native speakers is due to speaker variability, the locally recorded speakers not being professional newscasters. Fully-continuous, context dependent acoustic models were used, with a 25k-word vocabulary and word trigram language model for word hypotheses and a 52-word vocabulary and phone trigram language model for phoneme hypotheses. The same acoustic models were used in both cases.

4. WORD-DRIVEN DISCRIMINATION

In word-driven discrimination, word hypotheses are produced normally by the recognizer, in our case using the full LVCSR system described in Section 3.2. All hypotheses from a single speaker are bucketed into one document, and the set of documents from all speakers is used as training and testing data in the cross-validation scheme outlined in Section 2.2. Test "documents" can be restricted to a specific number of hypotheses to represent the number of utterances, or time, it would take for the classifier to make an accurate judgement about the nativeness of the speaker.

In addition to the word hypothesis classification, we ran a second set of experiments with words replaced by their parts of speech. This greatly reduces the number of unique word types used for classification, which was desirable because of the small number of training documents we had available. Using parts of speech to build the model also allows us to gain an understanding of the types of recognition errors that are common in non-native speech.

In evaluating classifier performance, four test conditions were defined:

Document source	word	POS
shared article	94	**100**
shared article (high-WER rec)	66	77
disjoint articles	**47**	**77**
train=d;test=s	56	95
train=s;test=d	56	83

Table 1. Classification accuracy of read speech. Baseline is 56%.

(a) train on shared article, test on shared article

(b) train on disjoint articles, test on disjoint articles

(c) train on shared article, test on disjoint articles

(d) train on disjoint articles, test on shared article

Table 1 shows results of classification on these four conditions. Classification results are always higher for the part-of-speech-tagged hypotheses than the word hypotheses. Although if one is training and testing on the same task, any discriminating feature in the data should be allowed to influence classification, we wished to establish the extent to which the difference in WER, as opposed to differences in the way the recognizer responds to non-native speech, contributed to classification performance. To accomplish this, we artificially increased the WER of the native speech to match that of the non-native speech by introducing white noise to the signal. This result is given in the second row of Table 1. Although the classification accuracy decreases somewhat, it is still significantly higher than chance, suggesting that there is indeed something special in the way non-native speech is being recognized that is independent of the word error rate.

Looking at the effect of train and test article mismatch, we see that when using words as features classification is only successful when the training and test hypotheses all originated from the same article. This is because of data sparsity. Looking at misrecognitions of words that appear a number of times in the data is quite telling; in an article about salmon, for example, it was easy to see that for native speakers *salmon* was most frequently misrecognized as *salmons*, while for non-native speakers it was misrecognized as *simon*, *someone*, and *some*, none of which occurred in native hypotheses. In the disjoint article case, however, where we do not have multiple examples of word misrecognitions, it was not possible to build a successful classifier from words. Using part-of-speech tags resulted in much stronger performance: 77% classification accuracy in the model built from parts of speech, as compared to 47% in the model built from words.

Document source	phone	phone class
shared article	100	86
disjoint articles	**92**	**80**
train=d;test=s	88	71
train=s;test=d	76	82

Table 2. Classification accuracy of read speech. Baseline is 58%.

5. PHONE-DRIVEN DISCRIMINATION

In phone-driven discrimination experiments, a phoneme string was produced by the recognizer instead of a word string. Classification was based, then, on how frequent specific phones were in the recognition hypotheses. As with the word and part-of-speech experiments, unigram and bigram tokens were considered as classification keys.

In addition to the phone hypotheses, a set of phone class hypotheses was produced in which each phone was replaced by a token for vowel (V) or consonant (C). This parallels the word-POS distinction, but as there are now only two classes, n-grams up to 5-grams were used for classification.

Results of phone-based discrimination are shown in Table 2. As with the word-based discrimination, using the phone identity, and not its class, is more accurate for condition-matched experiments. In the phone case, however, we do not need to be as concerned about overtraining on specific tokens, so there is not a compelling reason to use the poorer-performing phone classes.

The most interesting result is that when using phoneme hypotheses, training and testing on unique articles yields a classification accuracy of 92%. Table 3 shows the phone unigrams and bigrams that were most discriminative in this test case. Many of the phones indicative of native speech are ones that are known to be difficult for non-native speakers, particularly speakers of Japanese. When running phoneme

Phones		**Phone classes**	
Native	Non-native	Native	Non-native
dh	ih	CCC	V
th	hh	CC	VV
er	ao	CCCC	VCCV
axr	iy	C	VC
ax	ow	CCCCC	CVV
ax;th	aa	CCCCV	CV
ch	ih;ih	VCCCC	VVC
xn	ng	CVCCC	VCCVC
jh	ae	CCCVC	CVCCV
dh;ey	hh;ih	CCCV	CVVC

Table 3. Discriminative phone and phone class n-grams in phoneme hypotheses

recognition with no lexical model, these phonemes are simply not found in Japanese-accented speech. Instead, simple vowels like [a] and [i] are hypothesized with great frequency.

The consonant-vowel strings that are hypothesized, too, are not at all surprising when considering the two groups we are trying to discriminate. Frequent consonants and consonant clusters are clear indicators of native speech, while frequent vowels and CV-type syllables are indicators of Japanese-accented speech.

6. ACCENT-DEPENDENT RECOGNITION

With reliable accent discrimination, we can combine standard recognition with recently proposed techniques for adapting to non-native speech to run on-the-fly accent-dependent recognition [6, 1], e.g. Ideally, in such a system we would like to use disjoint sets of utterances for classifier training and testing, so we will use the phone-based classification, which achieved the best performance for disjoint articles. The algorithm for running accent-dependent recognition is as follows.

1. Generate a set of initial phone hypothesis using native context-dependent acoustic models, a lexicon with entries repesenting phonemes, and a language model built from phoneme distributions in the language model training corpus.
2. Pass the set of hypotheses through a classifier that has been trained on phoneme hypotheses of native and non-native speech
3. If the hypothesis is classified as native, re-recognize the speech with a word lexicon and a word language model
4. If the hypothesis is classified as non-native, re-recognize the speech with customized acoustic models, a word lexicon, and a word language model.

This process can be streamlined by generating word hypotheses in step 1 and classifying based on those hypotheses; if the speaker is judged to be native, the initial hypothesis will become the final hypothesis. Because the classification accuracy for word tokens is not as high as for phoneme tokens when testing on disjoint sentence sets, one could boost system performace either by using a common set for classification or biasing the classifier to prefer false negatives to false positives. We have found that falsely identifying native speakers as non-native is more harmful than falsely identifying non-native speakers as native; the mismatch between the native speech and the non-native acoustic models is severe.

Recognizer	WER		
	Native	**Non-native**	**Overall**
Baseline	21.6	58.1	42.2
Accent-dependent	22.5	45.1	35.6

Table 4. Recognizer performance with and without accent dependency

Table 4 shows how recognizer performance is improved when utterances identified as non-native by our classifier are re-recognized with customized acoustic models. Our non-native acoustic models were built by training the baseline Broadcast News models with 3 hours of accented acoustic data and interpolating this model set with a more robust set as described in [6].

In this experiment, our classifier produced one false positive, incorrectly identifying one native speaker as a non-native speaker. This is why the WER for the native speech increases when accent identification is applied. The overall WER, however, drops significantly.

7. CONCLUSION

In this paper, we have presented a fast and effective method for identifying non-native speech for LVCSR. We have found that Bayesian classification is extremely effective in detecting non-native utterances. We have also described an algorithm for integrating online classification with speech recognition which resulted in a 15% relative decrease in word error rate. The probabilistic properties of Bayesian models would allow this classification method to be used in combination with acoustic-feature-based identification for even greater accuracy.

8. REFERENCES

[1] Pascale Fung and Wai Kat Liu, "Fast Accent Identification and Accented Speech Recognition," in *Proc. ICASSP*, 1999.

[2] Carlos Teixeira, Isabel Trancoso, and António Serralheiro, "Accent identification," in *Proc. International Conference on Spoken Language Processing*, Philadelphia, 1996.

[3] Andrew Kachites McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," http://www.cs.cmu.edu/~mccallum/bow, 1996.

[4] Laura Mayfield Tomokiyo, "Linguistic Properties of Non-native Speech," in *Proc. ICASSP*, 2000.

[5] Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries, and Torsten Zeppenfeld, "The JanusRTk Switchboard/Callhome 1997 Evaluation System," in *Proc. the LVCSR Hub5-e Workshop*, 1997.

[6] Laura Mayfield Tomokiyo, "Lexical and Acoustic Modeling of Non-native Speech in LVCSR," in *Proc. ICSLP*, 2000.