HYPOTHESIS-DRIVEN ACCENT DISCRIMINATION

Laura Mayfield Tomokiyo

Carnegie Mellon University
laura@cs.cmu.edu

10.21437/Eurospeech.2001-26

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 15.6% 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 necessary if non-native modeling is to be fully integrated into an LVCSR system.

Prior work in accent discrimination has centered 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] used HMMs, but as traditional acoustic models in a six-way accent identification task. In their framework, competing models are trained for each native/spoken language pair. Acoustic-based approaches have a history of success in language identification (LID), where the language being spoken can be determined based on spectral [3] or prosodic [4] similarity to examples of language-tagged training speech, likelihood scores from competing monolingual recognition systems for word-level decoding of a full utterance [5], and phonotactics of a recognized phone sequence [6].

Language identification and accent discrimination, however, are fundamentally different tasks. In a language identification system, users are generally native or near-native speakers of the language they are using. Although speaker- and mode-dependent variation is widespread in native speech, the intuition that there is a consistency in native articulation of phones is substantiated by the effectiveness of the statistical models built to capture it. Non-native speakers, on the other hand, are each at a different point in the acquisition of the new phonological system. Particularly for lower proficiency levels, there is a large amount of inconsistency even within one speaker's articulation that cannot be traced to the speaker's native language [7]. This observation seems to conflict with the obvious agreement among native speakers on characteristics of certain foreign accents; it must be remembered, however, that native listeners' perception of accent is colored by their own phonological system and does not necessarily reflect the speaker's intent or the actual spectral properties of the articulatory event.

It was for this reason that we initially sought a discrimination method that would not depend on acoustic features. While accent discrimination cannot rely on the consistency in articulation that LID can, it has the benefit of knowing what language is being spoken, and what word sequences are characteristic of native speech.

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. The same method can also be applied to transcripts, if they are available, allowing non-native utterances to be identified in language model training data, for example. The method we describe is extremely fast and requires neither linguistic knowledge nor feature extraction. Hypotheses from an English recognizer trained on native speech are classified as native or non-native by a naive Bayes classifier that has been trained on examples of native and non-native speech hypotheses. If the speaker is found to be non-native during initial adaptation, the system switches to a customized acoustic model for optimal recognition accuracy.

2. BAYESIAN CLASSIFICATION

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.

2.1. Implementation and execution

The Rainbow statistical text classification package [8] was used for all classification experiments. Rainbow implements a naive Bayes classifier for text, with a number of specializations for text applications. Features are word n-tuples, with constituent n-tuples treated as independently occurring events.

The conventional usage of this type of package would be to classify text documents, such as newspaper articles, with respect to a property like topic or publication of origin. To frame accent identification as a text classification problem, each set of utterances from one training speaker, either hypotheses or transcripts depending on the experiment, 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. When testing on hypotheses, the model is trained on hypotheses; when testing on transcripts the model is trained on transcripts.

In building and testing the model, 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 trials. The baseline accuracy of this model is 56%, or the accuracy that would be obtained by always guessing the class best represented in the training data.

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. All articles are transcribed; because of reading errors and disfluencies the transcripts can be quite different from the original text. This data set is described in more detail in [9].

3.2. Recognizer

The JRTk speech toolkit [10] 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 18% for local native speakers and 58% for non-native speakers. It has been established that the lower performance for local native speakers is due to speaker variability. 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. Phoneme error rate is 52%. The same acoustic models were used in both cases; MLLR speaker adaptation is always applied.

4. WORD-BASED DISCRIMINATION

In word-based discrimination experiments, word hypotheses were produced offline for all speakers. All hypotheses from a single speaker's reading of an article 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 2.1. Training "documents" are used in their entirety to build models of native and non-native speech.

In addition to the baseline word hypothesis classification, two additional experiments were run to smooth the vocabulary used for classification. In the first, words were replaced by their parts of speech. This reduces dependency on, in particular, the nouns that occurred in each article. In the second experiment, only function words and extremely common words were considered for classification. We used the standard long SMART [11] list, which is ordinarily used to specify words to be *excluded* from the vocabulary; we used it for the opposite purpose because initial experiments had indicated that function words were highly discriminative. Both of these methods greatly reduce 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:

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

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

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

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

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. Restricting the vocabulary to stopwords greatly improves performance for matched-condition classification but not for the mixed conditions.

Looking at the effect of train and test article mismatch, we see that when using all words as features, classification is only successful when the training and test hypotheses all originated from the same article (condition A). This is a condition that would occur in an enrollment task, for example, when speakers "tune" the system to their own speech. In enrollment, all new users are typically asked to read from a specific text. Because the vocabulary is limited to the words that occur in this text, and all speakers are supposed to read the same words, training a very specific model to detect variation in recognition of those words is valid and desirable. 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. For identifying non-native speech, the hypothesized word *that* was the most discriminative token. The word *the* is often misrecognized as *that* in the non-native speech. This may be because of the speakers' tendency to produce the vowel with full quality, resembling an [a] instead of a schwa. Speakers also tended to articulate this familiar word very strongly and then pause in preparation for articulation of a possibly less familiar noun, resulting in a marked glottal stop that is often recognized as a [k] in unrestricted phoneme recognition, and a [t] when the lexical model constraints are added.

In the disjoint article case (condition B), 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. The most discriminative token in part-of-speech experiments was the plural noun for native speakers and the past-tense verb for non-native speakers. This was true for hypotheses in all conditions, including condition A, in which all speakers read the same article, so it reveals a tendency of the non-native speakers to read a singular noun when the noun in the text

Condition	Word		POS		Stopwords	
	trans	hypo	trans	hypo	trans	hypo
A	83	94	74	100	49	99
B	41	47	40	77	43	89
C	56	56	56	95	56	56
D	56	56	56	83	56	56

Table 1. Word-based classification accuracy of read speech. Conditions are defined in Section 4

was plural, rather than a preference on the part of native speakers for plural nouns. As for the important role the past-tense verb plays in identifying non-native speakers, it is our interpretation that the non-native speakers move less smoothly from word to word, and that epenthesized vowels and word fragments following verbs are taken by the recognizer to be a past tense ending.

Using the stopword list as the vocabulary creates a more effective classifier for the matched conditions (A and B) than the mixed conditions, even for the most difficult all-disjoint case, in which classification accuracy reaches 89%. For this condition, the most discriminative native word was *of* and the most discriminative non-native word was *to*. *Of* was very frequently misrecognized in the non-native speech, meaning that when it was recognized, the speaker was likely to have been native. As with the word *the* discussed above, non-native speakers had difficulty with the reduced quality of the vowel, resulting in the frequent recognition of *aux* for *of*. The word *aux* is not frequent and rarely used outside the context of an item borrowed from French; it should be dispreferred by the language model, but the combination of the high acoustic score and the fact that *aux* and *from* were assigned to the same semantic class contributed to this common misrecognition.

5. PHONE-BASED DISCRIMINATION

In phone-based 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 of length up to 5 were used for classification.

Results of phone-based discrimination are shown in Table 2. Because these recognition-based experiments, no figures for transcripts are given. As with the word-based discrimination, using the phone identity, and not its class, is more accurate for matched-condition 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 only condition in which phone-based classification outperforms word-based classification is condition B, where all speakers are reading different articles. This is also possibly the condition of most relevance for LVCSR. The difference in performance is not large; it is equivalent to misclassification of one speaker. Combining word and phone features does not improve classification accuracy, as can be seen from the column labeled *word+phone* in Table 2.

Condition	Phone	Phone class	Word+phone
A	100	86	100
B	92	80	86
C	88	71	41
D	76	82	47

Table 2. Phone-based classification accuracy of read speech

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 recognition with no lexical model, these phonemes are simply not found in the 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 [12, 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 word-based classification with the stopword vocabulary, which achieved strong performance for this condition. The algorithm for running accent-dependent recognition is as follows.

1. Run speaker adaptation normally using the adaptation articles and native acoustic models
2. Pass the set of adaptation hypotheses through a classifier that has been trained on word hypotheses of native and non-native speech
3. If the speaker is classified as native, continue with decoding
4. If the speaker is classified as non-native, switch to non-native acoustic models, re-adapt, and continue with decoding

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

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 [12]. The first column shows recognizer performance for the native test set using the baseline (native models) and accent-dependent systems. In this experiment, our classifier produced one false positive, incorrectly identifying one native speaker as a non-native speaker; that speaker therefore was recognized using the non-native models in the accent-dependent system. This is why the WER for the native speech increases when accent identification is applied. The second column shows recognizer performance for the non-native test set. The 58.1% baseline WER is the recognition performance of the native models on the non-native speakers. No non-native speakers were misclassified using the stopword-based naive Bayes method; the 45.1% WER for non-native speakers represents performance of the non-native models on the non-native test set. The third column shows performance of the recognizer for the combined native and non-native test set. The baseline system (recognizing all speakers with the native models) reaches 42.2%; when accent discrimination is used to switch between model sets, the word error drops to 35.6%, a 15.6% relative improvement in recognizer performance.

7. DISCUSSION

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. Classification accuracy with this method is *better* for hypotheses than for transcripts. We have also described an algorithm for integrating online classification with speech recognition which resulted in a 15.6% relative decrease in word error rate.

Results of our experiments show that different classifiers are effective in different training and testing contexts. For a situation such as enrollment in which all speakers will be reading from the same text, a classifier trained on parts of speech from word hypotheses offers the best performance, at 100% accuracy. This classifier also performs well when the training and test sources are mixed; that is, when the training speakers all read from the same text and the test speakers all read from disjoint texts, or vice versa. Both of these are reasonable scenarios; to minimize the training data collection effort or have speakers read from a certain phonetically balanced text researchers might prefer to have training speakers all read the same article where end users would all be expected to dictate different texts. The converse situation would be if the researchers have readings of disjoint texts available and do not wish to do fur data collection for classifier training, while the end users are all to read the same text, as is possible in a speaker verification application, for example.

The only condition for which the part-of-speech based classifier does not perform best is when all of the training and test articles are disjoint. In this case, the phone-based classifier is most accurate, followed closely by the stopword-based classifier. If the phoneme recognition step is too expensive for the application (regardless of the classification a second decoding pass would have to be run to obtain a word hypothesis), stopword-based classification can be substituted without a large degradation in accuracy.

8. FUTURE WORK

As a stand-alone method of classification, this approach is valuable when the recognizer is to be treated as a black box, as in a speech translation system, for example. Without accessing the acoustic features at all, researchers in machine translation can obtain an accurate classification of nativeness which can be used to trigger nativeness-dependent parsing and dialogue modeling, which we hope to explore.

The probabilistic properties of Bayesian models would also allow this classification method to be used in combination with acoustic-feature-based identification for even greater classification accuracy.

9. 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. ICSLP*, Philadelphia, PA, 1996.

[3] S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-independent, Text-independent Language Identification by HMM," in *Proc. ICSLP*, 1992, pp. 1011–1014.

[4] A.E. Thyme-Gobbel and S.E. Hutchins, "On Using Prosodic Cues in Automatic Language Identification," in *Proc. ICSLP*, 1996, pp. 1768–1772.

[5] Tanja Schultz and Alex Waibel, "LVCSR-based Language Identification," in *Proc. ICASSP*, 1996, pp. 781–784.

[6] L.F. Lamel and J.-L. Gauvain, "Cross-lingual Experiments with Phone Recognition," in *Proc. ICASSP*, 1993, pp. 507–510.

[7] E. Brière, "An investigation of phonological interference," *Language*, vol. 42, no. 4, pp. 768–796, 1966.

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

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

[10] 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.

[11] C. Buckley, "Implemetation of the Smart Information Retrieval Retrieval System," Tech. Rep. 85-686, Cornell University, 1985.

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