Accent Classification for Speech Recognition

Arlo Faria

International Computer Science Institute, Berkeley CA 94704, USA
arlo@icsi.berkeley.edu

Abstract. This work describes classification of speech from native and non-native speakers, enabling accent-dependent automatic speech recognition. In addition to the acoustic signal, lexical features from transcripts of the speech data can also provide significant evidence of a speaker's accent type. Subsets of the Fisher corpus, ranging over diverse accents, were used for these experiments. Relative to human-audited judgments, accent classifiers that exploited acoustic and lexical features achieved up to 84.5% classification accuracy. Compared to a system trained only on native speakers, using this classifier in a recognizer with accent-specific acoustic and language models resulted in 16.5% improvement for the non-native speakers, and a 7.2% improvement overall.

1 Introduction

Automatic speech recognition systems are highly susceptible to speaker variability. Statistical analysis reveals that – after gender – the principal component of this inter-speaker variation is accent [2]. Recognition models trained on one type of accent fare poorly when evaluated on a mismatched test condition. For this reason, most speech technology research is restricted to North American dialects of English, while the collected corpora mostly comprise native speakers.

With improving performance of speech recognizers and their expanding applications, the need to address non-native speakers has gained importance. Two recent speech corpora reflect the necessity of this research. The massive Fisher corpus[1] includes a considerable number of recruited subjects who speak English as a second language; meanwhile, the European Commission's AMI Project[2] is collecting data from meetings with many non-native English-speaking participants, as well as native speakers of non-American varieties of English.

To address the problems that non-native speakers present to speech recognizers, previous work has relied upon non-native accented training data. Adapting and retraining acoustic models from an accented corpus improved recognition of Japanese-accented English [8], and similarly with a Hispanic-English corpus [3]. Acoustic model adaptation can also be derived from a speaker's source language [5], data which is potentially more accessible. Alternatively, lexicon adaptation [9] can be utilized to reflect the phonology of non-native pronunciation.

[1] http://www.ldc.upenn.edu/Projects/EARS

[2] http://www.amiproject.org

This work presents an integrated accent-dependent speech recognition architecture that is analogous to gender-dependent systems. An accent classifier divided training data into native and non-native sets, from which recognition models were estimated; test data was similarly split and recognized with the corresponding accent-specific models.

We give careful consideration to the lexical aspects of non-native speech. Exploratory language modeling experiments suggested that the word structure of non-native language can be distinctly different from the native variety. Thus, in addition to the acoustic signals, text transcripts of the speech data were also used for accent classification.

To maximize the amount of data, non-native speakers were considered as a whole, rather than working with just one specific accent group. This treatment was partly justified by the better performance of a lexical classifier compared to an acoustic classifier. While non-native accents might sound quite different from each other, the words that these speakers generate tend to be characteristic of their non-native identity.

2 Data Preparation

These experiments were performed using a subset of the Fisher collection:

- **Speakers**: 948 speakers; 540 male, 408 female
- **Duration**: 158 hours = 948 speakers × 10-minute sides
- **Words**: 843 words per speaker, on average
- **Segments**: 90.5 segments per speaker, on average

The audio speech signals were recorded over 8 kHz telephone channels, and were accompanied by human-generated word-level reference transcripts. An acoustic speech segmentation tool automatically created segments without regard to sentence or phrase structure, although these segments were treated like sentences for language modeling purposes (i.e., affixed with the boundary tags `<s>` and `</s>`).

Self-reported participant information was gathered to describe speaker demographics, and trained human auditors rated each speaker's accent as *American* or *Other*. For these experiments, the non-native speakers were those whose native language was not reported as English and whose accents were audited as *Other*. The set of native speakers reported English as a native language and had accents audited as *American*[3]; a subset was selected to match the size and gender proportions of the non-native set. Normalizing the amount of data per speaker, we used just one 10-minute conversation side for each speaker.

Table 1 gives the composition of the native and non-native accented sets. Native speakers are grouped by place of birth, with many locally recruited participants originating from the American Northeast. The non-native portion is categorized by speakers' self-reported native languages. These groupings are only for description of the data sets; in this paper, only the native versus non-native distinction is considered.

[3] The Fisher collection explicitly excluded British speakers from participation.

Accent Type	Speakers	Male	Female
Non-native	474	270	204
Indian	116	85	31
Chinese	102	50	52
Russian	61	23	38
Spanish	60	36	24
German	26	13	13
French	20	7	13
Other	89	56	33

Accent Type	Speakers	Male	Female
American English	474	270	204
Pennsylvania	60	39	21
New York	56	36	20
California	53	32	21
Texas	21	11	10
New Jersey	18	10	8
Ohio	19	11	8
Other	247	131	116

Table 1. Fisher corpus demographics

In Table 1, the non-native accents are grouped as follows:

- **Indian** is primarily Hindi. Also: Tamil, Farsi, Urdu, Telegu, Bengali, Marathi, Gujarati, Malayalam, Kannada, Punjabi, and Sindhi.
- **Chinese** includes Mandarin and Cantonese.
- **Russian** comprises Russian, Bulgarian, Hungarian, Polish, Czech, Armenian, Serbian, Croatian, Bosnian, Slovak, and Latvian.
- **Spanish** speakers are mainly Hispanic and Latin Americans. Also included are the West Iberian languages: Portuguese and Galego.
- **German** comprises mostly Germanic languages: German, Danish, Dutch, Swedish, and Afrikaans.
- **French** speakers are from France, Canada, and Switzerland.
- **Other** languages (with four or more speakers): Arabic, Turkish, Korean, Creole, Yoruba, Romanian, Japanese, Hebrew, Greek.

Test and training sets of 100 and 374 speakers, respectively, were selected from the native and non-native sets above, ensuring that the composition of each subset reflected the proportions given in Table 1.

3 Accent Classification

3.1 Acoustic GMM Classifier

Given accent-specific acoustic models λ_a that assign probability to acoustic observations X, we can invoke the maximum likelihood criterion to determine the accent classification $\hat{a}$:

$$\hat{a} = \arg\max_a \mathrm{P}(X|\lambda_a)$$

Ideally, the acoustic models in this computation would be a set of accent-specific phone HMMs used for recognition; however, then it is usually necessary to align X to a phone sequence determined in an earlier decoding pass. A more efficient solution implements λ_a as a Gaussian Mixture Model: a global distribution of speech frames, independent of sequence. This application of GMMs is fairly standard in other speech classification tasks such as speech detection, gender classification, speaker identification, and warp factor selection for vocal tract length normalization.

	Native train	Non-native train
Native test	143	153
	1.78%	2.31%
Non-native test	146	135
	1.53%	1.68%

Table 2. Perplexity and out-of-vocabulary rate.

	Native train	Non-native train
Native test	14.09	14.31
Non-native test	14.28	14.05

Table 3. Perplexity of a POS sequence model.

Accent-specific acoustic GMMs were built from the native and non-native training data; each was a mixture 256 Gaussians trained for 10 iterations of EM. Acoustic observations were standard ASR features: 12 mel-frequency cepstral coefficients and energy, plus their first and second order derivatives. Features were transformed with speaker-based cepstral mean/variance normalization, and also with vocal tract length normalization to counteract the models' gender independence.

3.2 Lexical SVM Classifiers

Non-native accented speakers of English are often distinguished by acoustic divergence from the standard pronunciation of native speakers. Beyond the phonetics and phonology, however, non-native speakers generally have a weaker command of the language and consequently produce sequences of words that a native speaker would be less likely to utter. The motivation for using lexical features for accent classification is based upon this hypothesis that non-native speakers produce word sequences that are fundamentally different from the language produced by native speakers. Before attempting to work with lexical features, however, it would be reassuring to test this hypothesis with some simple language modeling experiments.

Language models were built from each of the accented training sets (about 300K words each), and the perplexity was calculated for each of the accented test sets (about 100K words). These open-vocabulary trigram models were smoothed using Chen and Goodman's modified Kneser-Ney discounting scheme, implemented in the SRI Language Modeling Toolkit [6]. Table 2 demonstrates the results of training and testing language models on various combinations of the native and non-native sets. There is a clear correlation between matched accent conditions and lower perplexity. Because non-native word sequences are better predicted by training on non-native speakers, this suggests that there is a distribution of characteristic words and phrases that differs from the native

Feature type	Classifier type	Classification accuracy
Acoustic MFCCs	GMM	69.5%
Word Unigrams	SVM	74.5%
Word Bigrams	SVM	75.0%
Word Trigrams	SVM	76.5%
POS Unigrams	SVM	68.5%
POS Bigrams	SVM	70.5%
POS Trigrams	SVM	72.5%
All Lexical	Interpolated	77.5%
All Lexical + Acoustic	Interpolated	82.0%
Word Trigram + Acoustic	Interpolated	81.5%

Table 4. Accuracy of accent classification.

set's. Additionally, the non-native test set had a significantly lower out-of-vocabulary rate, reflecting the understandably smaller vocabulary size of speakers who have had less exposure to the English language.
Table 3 provides more evidence supporting the hypothesis that language generated by non-native speakers is different. A rule-based tagger trained from WSJ data [1] assigned Penn Treebank part-of-speech tags to all the data, allowing the estimation of a part-of-speech trigram model. Again there is a correlation between matched accent conditions and better predictability of tag sequences. This might be attributed to a preference for certain syntactic forms and tenses. Or it is possibly related to grammatical errors committed by language learners: auxiliary and function words tend to be misused; if POS tags convey some morphological information, it would also be possible to detect errors of agreement.
Given these results, two kinds of word-based features were investigated for accent classification:

- Word n-grams. The distribution of words and word sequences is different for each accent group, so n-gram counts could be good features for categorization of speaker accents given their text transcripts.
- POS n-grams. There are probably some sequences of part-of-speech tags that native speakers rarely produce, but are more commonly misused by non-native speakers.

The integral counts of these n-grams were provided as input features for text categorization with Support Vector Machines, using the SVM-Lite toolkit [4] with a linear kernel. The training algorithm was presented with 748 accent-labeled data points, one for each conversation side.

3.3 Comparison of Accent Classifiers

For the accent classifiers described, performance on the test set of 200 speakers was evaluated by comparing to the reference judgments made by human auditors of the Fisher corpus. As a baseline, the prior probability of each accent (native or non-native) was exactly 50%.

Feature type	Classifier type	Classification accuracy
Word Trigrams with reference transcripts	SVM	76.5%
Word Trigrams with recognized hypotheses	SVM	79.0%
Word Trigrams (ref trans) + Acoustic	Interpolated	81.5%
Word Trigrams (rec hyps) + Acoustic	Interpolated	84.5%

Table 5. Lexical features from reference and recognized transcripts.

Two types of classifiers were used: a maximum-likelihood GMM for the acoustic features, and a SVM for the lexical features. Neither classifier returned normalized probabilities, so the combination of these scores was accomplished by linear interpolation (summation of weighted scores), tuned with a grid search over the mixing weights[4]

The performance of all the classifiers is given in Table 4, where the optimal combination of all features achieved 82% accuracy; combining the SVM score for word trigrams with the acoustic GMM score was sufficient for 81.5% accuracy, and for simplicity this was the scheme chosen for the experiments in the next section.

In all experiments described thus far, lexical features were extracted from human-annotated reference transcripts. In the next section, we will describe ASR architectures that use accent classifiers with lexical features extracted from 1st-pass recognition hypotheses. Despite the high word error rate, the accent classification accuracy actually improves, as shown in Table 5. This convenient result suggests that the errors made by the recognizer are perhaps also correlated to a speaker's accent.

4 Accent-dependent Speech Recognition

An accent classification system is not very practical on its own, and in this project its intended application is to pre-process the data used in an accent-dependent speech recognition system. Using accent-specific models can greatly improve recognition performance, but relies upon a good accent classifier to appropriately select which models to apply.

Several ASR systems were built and tested with SRI's DECIPHER [7]. The resulting performance was suboptimal because many compromises were made to allow for rapid training and testing of the systems, as well as to provide a carefully controlled experiment. In particular, the gender-independent acoustic models (genonic HMMs) were trained on a relatively small amount of data: about 60 hours of the quickly annotated Fisher corpus, rather than hundreds of hours of precisely transcribed speech. The language models were exclusively trained on the small subsets of the Fisher data: bigrams can be rather sparse with only 300K words of training text. Also, there was no speaker adaptation of acoustic

[4] This was a "cheating" experiment: tuned on the test set. However, there was not a sharp peak at the optimal interpolation weight.

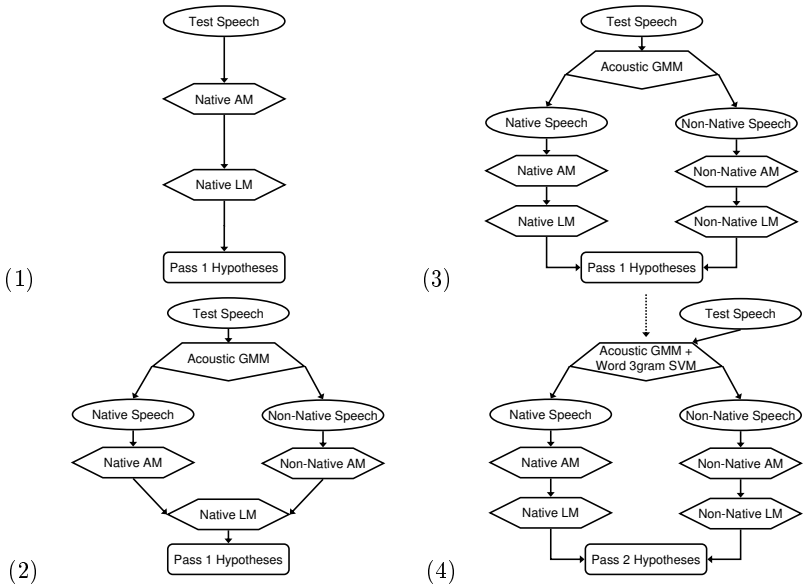

Fig. 1. Four types of system architectures: (1) Baseline system using native models; (2) Accent-specific acoustic models; (3) Accent-specific acoustic and language models; (4) Two-pass system using lexical and acoustic features for accent classification.

models – only VTLN in the front-end feature extraction. These optimizations allowed very fast run-time performance, as the recognizer processed speech data in less than 3x real-time on a 2.4GHz Pentium machine.
Figure 1 depicts various accent-dependent architectures. In System (2), an acoustic GMM classifier selects the accent-specific acoustic models. System (3) is similar, but the language models are also accent-specific. The first-pass recognition hypotheses from System (3) are utilized in System (4) to classify accents using acoustic and lexical features.

4.1 Results of recognition experiments

We first consider the separation of native and non-native speakers according to the judgments of the human auditors. Table 6 describes the performance of accent-dependent recognizers when models are matched and mis-matched to the test accents. The first column represents a system trained only on native speakers, System (1). The rightmost column represents a gold-standard system, if it could use the human auditors to select which accent-specific recognition models to employ.
Results of the recognition experiments are summarized in Table 7, demonstrating how an automatic speech recognition system can improve performance by identifying non-native speakers with lexical information, as well as acoustic, and recognizing those speakers with non-native models. These results again support the hypothesis that non-native speakers differ in the lexical aspects of their language use.

	Native models	Non-native models	Accent-matched models
Native Test	**50.72**	59.30	**50.72**
Non-native Test	64.40	**52.79**	**52.79**
Overall	57.20	56.22	**51.70**

Table 6. Combinations of accent-specific models: Word Error Rate %

System	Native	Non-native	Overall
(1)	50.72	64.40	57.20
(2)	53.76	53.45 (-17.0%)	53.62 (-6.3%)
(3)	53.73	53.32 (-17.2%)	53.55 (-6.4%)
(4)	52.64	53.75 (-16.5%)	53.08 (-7.2%)
Gold Standard	50.72	52.79 (-18.0%)	51.70 (-9.6%)

Table 7. Results of speech recognition experiments: Word Error Rate %.

In retrospect, it would have been informative to compare these results to an accent-independent system with models trained on all the data, not just the native set. Models trained on twice as much data would be less sparse; however, combining the accents would also make the distributions less sharp. This is a possibility for future experimentation.

5 Conclusion

This work described a series of experiments using subsets of native and non-native speakers drawn from the Fisher corpus. An investigation of the word and part-of-speech sequence models gave evidence that speaker accents are more than simply acoustic differences. Lexical features proved useful for accent classification, even when extracted from relatively poor recognition hypotheses. Lastly, accent classifiers were integrated into an accent-dependent speech recognition architecture which significantly outperformed a system trained only on native speakers.

Similar to physiological factors such as gender, accents contribute to the general problem of speaker variability. As speech recognition systems evolve to address these challenges, the utility of the technology increases and it becomes more accessible to diverse populations. In this global perspective, modern speech recognizers must be designed to perform for all kinds of accents, and not exclusively native speakers.

Acknowledgements

Thanks to Barbara Peskin and Dan Klein for advising this work; Andreas Stolcke, Kofi Boakye, and Andy Hatch provided invaluable technical support with the implementation; also, Javier Macías-Guarasa's accent adaptation work was a strong inspiration.
This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811).

References

1. E. Brill. A report of recent progress in transformation-based error-driven learning. In *AAAI*, 1994.
2. C. Huang, E. Chang, and T. Chen. Accent issues in large vocabulary continuous speech recognition. Technical Report MSR-TR-2001-69, Microsoft Research China, Beijing, China, 2001.
3. A. Ikeno et al. Issues in recognition of Spanish-accented spontaneous english. In *Proceedings of IEEE/ISCA Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.
4. T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proc. of European Conference on Machine Learning*, 1998.
5. L. W. Kat and P. Fung. MLLR-based accent model adaptation without accented data. In *Proceedings of ICSLP*, 2000.
6. A. Stolcke. SRILM – an extensible language modeling toolkit. In *Intl. Conf. on Spoken Language Processing*, 2002.
7. A. Stolcke et al. The SRI March 2000 Hub-5 conversational speech transcription system. In *Proc. NIST Speech Transcription Workshop*, University of Maryland, May 2000.
8. L. M. Tomokiyo. *Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in LVCSR*. PhD thesis, Carnegie Mellon University, 2001.
9. W. Ward et al. Lexicon adaptation for LVCSR: Speaker idiosyncracies, non-native speakers, and pronunciation choice. In *Proceedings of the PMLA Workshop*, Estes Park, Colorado, 2002.