AN EMPIRICAL STUDY OF AUTOMATIC ACCENT CLASSIFICATION

Ghinwa Choueiter

Massachusetts Institute of Technology
32 Vassar Street
Cambridge, MA 02139
ghinwa@mit.edu

Geoffrey Zweig, Patrick Nguyen

Microsoft Research
One Microsoft Way
Redmond, WA 98052
{gzweig,panguyen}@microsoft.com

ABSTRACT

This paper extends language identification (LID) techniques to a large scale accent classification task: 23-way classification of foreign-accented English. We find that a purely acoustic approach based on a combination of heteroscedastic linear discriminant analysis (HLDA) and maximum mutual information (MMI) training is very effective. In contrast to LID tasks, methods based on parallel language models prove much less effective. We focus on the Oregon Graduate Institute Foreign-Accented English dataset, and obtain a detection rate of 32%, which to our knowledge is the best reported result for 23-way accent classification.

Index Terms— Accent classifier, GMM, MMI, Gaussian tokenization, language identification.

1. INTRODUCTION

Accent classification is the task of automatically detecting the accent of a foreign speaker from a spoken utterance. In this research, we target accent classification in foreign-accented English with the aim of embedding such a classifier within Voice-Rate, an experimental dialogue system [1]. The Voice-Rate system provides product ratings over cell-phones to consumers via a toll-free number, and accent classification would enhance it by providing the necessary information to perform consumer profile adaptation and eventually targeted advertising based on consumer demographics.

There has been little past research in the area of accent classification. In particular, most of the previous work in the field involves only two- to four-way classification. Deshpande and colleagues used the second and third formants and Gaussian Mixture Models (GMMs) to achieve a detection rate of 86% on American versus Indian American accent classification [2]. Gray and Hansen used pitch and formant contours and voice onset time and Stochastic Trajectory Models (STMs) to distinguish between American, Chinese, and Turkish accents [3]. They achieved detection rates of 90.4% and 52.1% on read and spontaneous speech respectively. In [4], STMs were evaluated against GMMs and HMMs on 4-way accent classification, where the accents considered were Chinese, Thai, Turkish, and American. The classification rates obtained were 40.1% with GMMs, 41.3% with HMMs, and 41.9% with STMs. To the authors' knowledge, there has only been one previous work [5] that evaluates accent classification on the same dataset used in this research. This previous work, which has not been formally published, reports detection rates of 73% and 58.9% for German versus Spanish classification using GMMs and naïve Bayes classification respectively. Further detection rates of 36.2%, 17.7%, and 13.2% are reported for 4-, 13-, and 23-way naïve Bayes accent classification.

Whereas accent detection is relatively unresearched, there has been a very significant amount of previous work in language identification (LID) [6, 7, 8, 9, 10, 11], and the main contribution of this paper is the extension of these methods to accent detection. Previous work on LID falls into one of three categories. In the first, language classification is performed using acoustic scores typically obtained using GMMs or phone recognizers [6, 7, 8, 9, 10]. In the second, the LID classification score is derived from a language model (LM), which captures the statistics of either phones generated by a phone recognizer or gaussian tokens corresponding to the gaussians with the highest likelihood in each time frame [6, 7, 8, 10]. Finally, in the third category, languages are modeled using vectors of phone statistics and are detected using text classification techniques [11].

In the rest of this paper, the corpus and the baseline are described in Sections 2 and 3 respectively. MMI training is evaluated in Section 4, and Gaussian Tokenization in Section 5. Section 6 concludes with a summary.

2. DATA

The corpus used in this research is the CSLU Foreign-Accented English (FAE) dataset [12]. The corpus consists of 4925 telephone-quality utterances spoken by native speakers of 23 languages. In no event is English a speaker's native language. Most of the utterances are 20 seconds in length, and none are phonetically transcribed. The Train, Development, and Test sets were created by randomly sampling and splitting the orig-

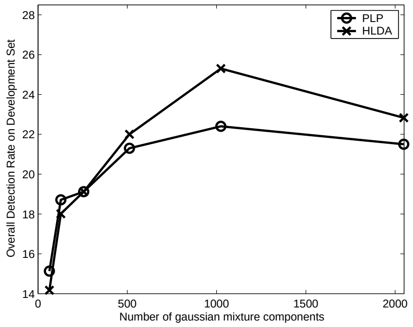

Fig. 1. The accent detection rate of the baseline as a function of GMM order, with HLDA. The detection rate is evaluated on the Development set.

inal corpus into a (70%,15%,15%) configuration. The 23 accents recorded in the CSLU FAE corpus are Arabic, Brazilian, Portuguese, Cantonese, Czech, Farsi, French, German, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Mandarin, Malay, Polish, Iberian, Portuguese, Russian, Swedish, Spanish, Swahili, Tamil, and Vietnamese.

3. THE BASELINE

For the baseline, a GMM is built for each of the 23 accents. A 52-dimensional acoustic observation is obtained from a 13-dimensional plp-based vector concatenated with its first, second, and third derivatives. Mean and variance normalization is performed on the acoustic feature vector prior to training the GMMs. The GMMs are initialized using k-means and trained with the EM algorithm. Accent classification is performed by selecting the model with the highest log-posterior. Uniform priors over the accents were used. In an initial set of experiements, we observed that uniform and non-uniform accent priors yielded identical detection rates.

Dimensionality reduction is also investigated with the feature dimension decreased from 52 to 39 using heteroscedastic linear discriminant analysis (HLDA) [13]. Figure 1 plots the detection rate of the baseline accent classifier, before and after applying HLDA, as a function of the GMM order which varies from 64 to 2048. We remark that (1) the baseline system benefits from HLDA for GMM orders larger than 256, (2) the classifier detection rate exhibits initial improvement as the GMM order is increased, peaks at 1024, and then deteriorates due to overtraining, and (3) our best detection rate for the baseline is 25.3% after applying HLDA (22.4% before HLDA).

4. MMI TRAINING

In the previous section, we described the baseline trained with the Maximum Likelihood (ML) criterion, which optimizes the log-likelihood of the training data. In this section, we investigate discriminative training with the Maximum Mutual Information (MMI) criterion [14, 15], which, when the language model is fixed, optimizes the log-posterior of the correct labels in the training data. Our implementation follows that of [16], using a constant value for E.

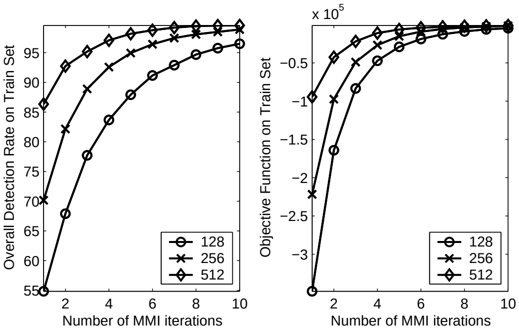

Fig. 2. The detection rate (left) and the MMI objective function (right) on the Train set as a function of MMI iterations for GMM orders 128, 256, and 512.

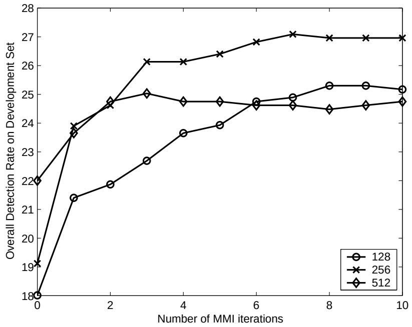

Fig. 3. Detection rate on the Development set as a function of MMI iterations for GMM orders 128, 256, and 512.

4.1. MMI Results

In this section, we report results for MMI training, and, in particular, we look at the effects of the number of MMI iterations, the GMM order, as well as the value of E, the global constant used in the computation of D.

MMI iterations and GMM order: The MMI training algorithm is initialized with an ML-trained accent classifier described in Section 3, and ten MMI update iterations are performed. GMMs of orders 128, 256, and 512 are investigated. First, we depict the performance of the GMM-based accent classifiers on the Train set as a function of MMI iterations. Results are reported in Figure 2 for all three GMM orders and for $E = 5$, and performance is evaluated both in terms of overall detection rate as well as MMI objective function. The results on the Train set exhibit the expected behavior where

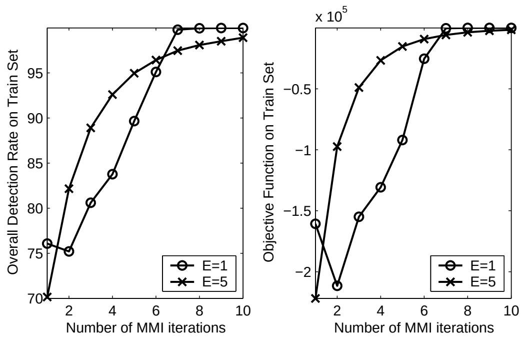

Fig. 4. Detection rate (left) and the MMI objective function (right) on the Train set as a function of MMI iterations for GMM of order 256 and $E = 1, 5$.

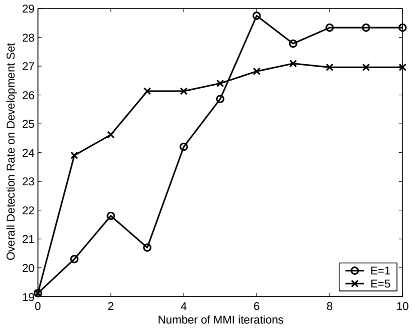

Fig. 5. Detection rate on the Development set as a function of MMI iterations for GMM of order 256 and $E = 1, 5$.

both detection rate and objective function improve steadily and then level off as more MMI iterations are performed.

Next, the overall detection rate of the accent classifier is reported on the Development set with $E = 5$ in Figure 3. The detection rates for orders 128 and 256 increase steadily with MMI iterations, however those for 512 indicate that the models are overtrained. For this reason MMI training is not performed for orders beyond 512. Following ten MMI iterations for order 256, a detection rate of 27% has been achieved.

MMI iterations and E: While Figure 2 shows steady improvement across iterations, we have found that for this particular task, the algorithm is quite sensitive to how E is set. Figure 4 illustrates the detection rate and objective function evaluated on the Train set for GMM order 256 and $E = 1, 5$. The results are consistent with the findings in [16], where the larger E, the larger D tends to be, and the more stable yet slower the MMI training becomes. The effect of E is evaluated on the Development set in Figure 5. Interestingly, the results for $E = 1$, though exhibiting a less steady ascent than $E = 5$, give a higher detection rate of 28.2% (27% for $E = 5$) after ten MMI iterations.

The performance of the accent classifier following ten MMI iterations is illustrated in Figure 6. Finally, the performance of the MMI-trained accent classifier is evaluated on the Test set itself, and a detection rate of 32% is obtained with GMM order 256 and $E = 1$.

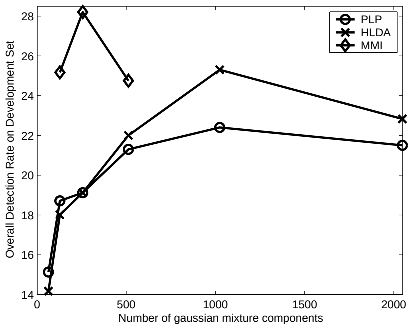

Fig. 6. Accent detection rate evaluated on the Development set, as a function of GMM order for the baseline GMM model, improved with HLDA, and MMI cummulatively.

5. GAUSSIAN TOKENIZATION

In this section, we briefly describe the Gaussian Tokenization (GT) approach that has been previously proposed for LID, and refer the reader to [6, 7] for more detail. In this research, a gaussian tokenizer is a GMM that generates a sequence of indices for an utterance, where each index corresponds to the mixture component with the highest likelihood in a time frame. The index sequences are then used to train index language models (LM) for each accent. In accent independent (AI) tokenization, a single set of gaussians is used across all accents, while in accent dependent (AD) tokenization, an accent specific GMM is used instead. The motivation behind this approach is that a tokenizer would generate sequences that exhibit different patterns for each accent, and that the statistics of these patterns could then be captured using an LM. Figure 7 illustrates the training setup using an accent-dependent GT. During decoding, illustrated in Figure 8, an utterance is presented to each AD GT, and each index sequence is fed to the corresponding AD LM. The LM which gives the lowest perplexity is selected.

Our experiments investigate both accent-dependent (AD) and independent (AI) GTs, where in the former case, 23 GMMs are trained with accent-specific data, and in the latter a single GMM is trained with all the data. We implement index LMs as *n*-grams of orders 2 to 5. The same data (Train set) is used to train the Gaussians and generate the index sequences. The results shown in Table 1 indicate that the AD approach somewhat outperforms the AI approach, but that overall, Gaussian Tokenization does not have the same success with accent classification as with LID. We note, however, that this is with ML trained LMs, and the results might improve significantly with discriminative LM training.

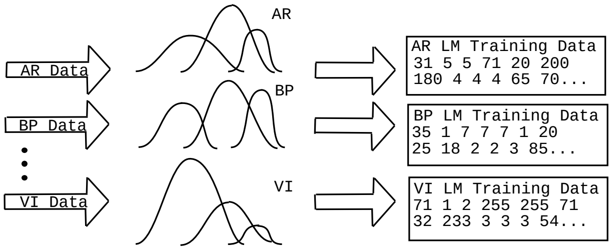

Fig. 7. Illustration of the training setup using an accent-dependent Gaussian Tokenizer.

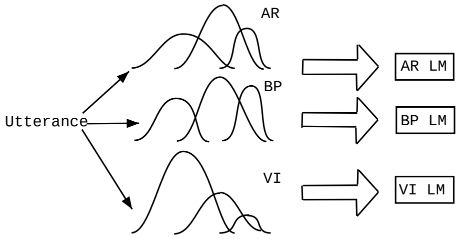

Fig. 8. Illustration of the decoding setup using an accent-dependent Gaussian Tokenizer.

LM scores obtained with the MMI-trained GT are interpolated with our best acoustic scores, improving the detection rate from 28.2% (c.f. Section 4.1) to 28.8% on the Development set, and from 32% to 32.7% on the Test set.

6. SUMMARY AND DISCUSSION

In this research, several approaches to accent classification have been presented and evaluated on the 23 accents in the CSLU FAE corpus. The results for the various methods are summarized in Table 2 for the Development and Test sets.

We find that acoustic-only methods are quite effective for accent classification, and that in contrast to typical LID systems, we see little improvement from incorporating language models scores based on subword symbol sequences.

GT Type	Detection Rate	GT Type	Detection Rate
128 2-gram(AD)	9.76%	256 MMI 3-gram(AD)	15.13%
128 3-gram(AD)	12.37%	512 2-gram(AD)	13.60%
128 4-gram(AD)	12.51%	512 3-gram(AD)	13.40%
128 5-gram(AD)	12.37%	1024 2-gram(AD)	13.20%
256 2-gram(AD)	13.75%	1024 3-gram(AD)	12.92%
256 3-gram(AD)	**16.23**%	1024 2-gram(AI)	11.14%
256 MMI 2-gram(AD)	12.65%	1024 3-gram(AI)	10.17%

Table 1. Performance of different types of GTs on the Development set. The description of the GT type includes the GMM order, the LM order, and whether the GT is accent dependent or independent.

Method	Development Set	Test Set
ML+HLDA	25.3%	23.1%
MMI+HLDA	28.2%	32.0%
GT+MMI+HLDA	28.8%	**32.7**%

Table 2. Summary of improvement in detection rates over the GMM baseline after cumulative application of HLDA, MMI, and GT.

7. REFERENCES

[1] G. Zweig, P. Nguyen, Y. C. Ju, Y. Wang, D. Yu, and A. Acero, “The voice-rate dialog system for consumer ratings,” in *Proc. Interspeech’07*, Antwerp, Belgium, 2007.

[2] S. Deshpande, S. Chikkekur, and V. Govindaraju, “Accent classification in speech,” in *Proc. Automatic Identification Advanced Technologies Workshop*, 1005, pp. 139–143.

[3] S. Gray and J. H. L. Hansen, “An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system,” in *Proc. Automatic Speech Recognition and Understanding Workshop*, San Juan, Perto Rico, 2005, pp. 35–40.

[4] J. H. L. Hansen P. Angkititrakul, “Advances in phone-based modeling for automatic accent classification,” *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, pp. 634–646, March 2006.

[5] J. Macias-Guarasa, “Acoustic adaptation and accent identification in the ICSI MR and FAE corpora,” in *ICSI Meeting slides*, 2003.

[6] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., “Approaches to language identification using gaussian mixture models and shifted delta cepstral features,” in *Proc. ICSLP’02*, Denver, Colorado, 2002, pp. 89–92.

[7] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller Jr., “Language identification using gaussian mixture model tokenization,” in *Proc. ICASSP’02*, Orlando, Florida, 2002, pp. 757–760.

[8] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, “Acoustic, phonetic, and discriminative approaches to automatic language identification,” in *Proc. Eurospeech’03*, Geneva, Switzerland, 2003, pp. 1345–1348.

[9] L. F. Lamel and J. L. Gauvain, “Language identification using phone-based acoustic likelihoods,” in *Proc. ICASSP’94*, Adelaide, Australia, 1994, pp. 293–296.

[10] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. Speech and Audio Proc.*, vol. 4, pp. 31–44, January 1996.

[11] H. Li, B. Ma, and C. Lee, “A vector space modeling approach to spoken language identification,” *IEEE Trans. Speech and Audio Proc.*, vol. 15, pp. 271–284, January 2007.

[12] “CSLU foreign-accented english corpus,” http://www.cslu.ogi.edu/corpora/fae/.

[13] N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. thesis, Johns Hopkins University, Baltimore, Maryland, 1997.

[14] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, “Maximum mutual information estimation of hidden markov model parameters for speech recognition,” in *Proc. ICASSP’86*, Tokyo, Japan, 1986, pp. 49–52.

[15] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem.*, Ph.D. thesis, McGill University, Montreal, Canada, 1991.

[16] P. C. Woodland and D. Povey, “Large scale discriminative training for speech recognition,” in *ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium*, Paris, 2000, pp. 7–16.