Phoneme-less Hierarchical Accent Classification[1]

Xiaofan Lin, Steven Simske
Imaging Systems Laboratory
HP Laboratories Palo Alto

HPL-2004-166
October 4, 2004*

E-mail: {xiaofan.lin, steven.simske}@hp.com

speech recognition, accent classification, interactive voice response system

This paper introduces a novel accent classification method. Compared with existing methods, it has two unique features. First, it does not explicitly utilize phoneme information. Second, it is built on top of the gender classification. We have tested the proposed algorithm on datasets that are completely independent of training data. The accuracy of distinguishing American accent and British accent is 83%. We have also compared the accent classification with the gender classification in terms of accuracy and the saturation behavior with respect to length of utterance.

* Internal Accession Date Only

[1] To be presented at and published in the 38th Annual Asilomar Conference on Signals, Systems, and Computers, 7-10 November 2004, Pacific Grove, CA

Approved for External Publication

Phoneme-less Hierarchical Accent Classification

Xiaofan Lin and Steven Simske

Hewlett-Packard Labs, 1501 Page Mill Road, MS 1203
Palo Alto, CA 94304, USA
Email: {xiaofan.lin, steven.simske} @hp.com

***Abstract*-This paper introduces a novel accent classification method. Compared with existing methods, it has two unique features. First, it does not explicitly utilize phoneme information. Second, it is built on top of the gender classification. We have tested the proposed algorithm on datasets that are completely independent of training data. The accuracy of distinguishing American accent and British accent is 83%. We have also compared the accent classification with the gender classification in terms of accuracy and the saturation behavior with respect to length of utterance.**

I. INTRODUCTION

Accent classification is important in speech recognition and speech-controlled applications. With accurate classification, we can train a speech recognizer for each accent, and then expect to have a higher recognition rate than with a system that does not distinguish accent. In addition, smart Interactive Voice Response (IVR) systems can be directly built on the basis of accent classification. For example, when a customer calls in, the agent that can best understand the detected accent will be assigned to serve the customer. In this way, both the customer satisfaction and the efficiency can be improved.

Most existing work on accent classification employs phoneme information. Miller and Trischitta used the average cepstral vectors of phonemes to distinguish different regional dialects in US [1]. Angkititrakul and Hensen introduced the Stochastic Trajectory Model (STM) for each phoneme to classify different foreign accents of English [4]. Fung and Wai Kat proposed a phoneme-class HMM for fast access identification, in which one HMM is shared by several phonemes within the same phoneme category [5]. Lincoln, et al, described a phonotactic model for the classification of the accent [2].

However, there are a couple of drawbacks associated with the use of phonemes. Phoneme classification itself is a difficult task and requires training for different languages. The pure phoneme recognition rate (without the knowledge of grammar or vocabulary) is only about 60% [12]. In addition, previous work has treated accent classification as an isolated task and did not integrate it with other speech metadata extraction tasks such as gender classification [6]. In this paper, we introduce a hierarchical accent classification without explicitly using phoneme information. To minimize the bias associated with a given dataset's acoustic characteristics, we have conducted extensive tests across independent datasets. Section II introduces the proposed phoneme-less hierarchical accent classification algorithm. Section III discusses the choice of speech databases. Section IV shows the experimental results. Section V introduces an application system based on the accent classification algorithm. Section VI gives a summary and draws conclusions.

II. PHNOEME-LESS HIERARCHICAL ACCENT CLASSIFICATION

The proposed accent classification algorithm has two unique features. First, it employs a hierarchical classification scheme. Second, it does not explicitly use phoneme information. This section will describe the two features in more details.

A. Hierarchical Classification

In the training stage, we have trained two models, one for the accents of male, and the other for the accents of female. In the testing stage, the first step is to recognize the speaker's gender. The method proposed in our previous work [6] is used to detect the gender. The accent models are then selected based on the detected gender. Then the accent classifier is invoked using the selected accent models. The overall workflow is shown in Fig. 1.

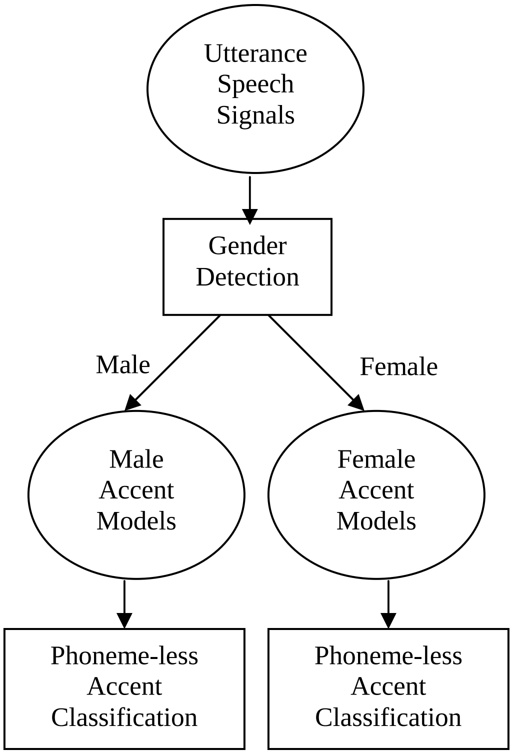

Fig. 1. Hierarchical accent classification

B. Phoneme-less Accent Classification

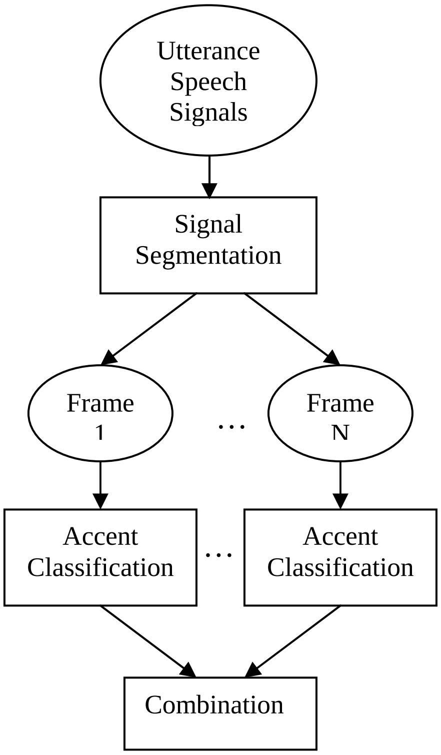

Fig. 2. Phoneme-less accent classification

The actual accent classification does not utilize any phoneme information and operates as shown in Fig. 2. Each utterance is first evenly segmented at an interval of 10 milliseconds. The size of each frame is 25.6 milliseconds. Then the 13 MFCC features are extracted from each frame using the algorithm implemented in Carnegie Mellon University's Sphinx 2.0 Speech Recognition System [7]. The first component only reflects the energy, so the remaining 12 features are used in the accent classification. Cepstral Mean Subtraction (CMS) is applied to the raw feature to reduce the influence of channel mismatch [13]. The classifier is based on Gaussian Mixture Model (GMM). Each accent uses 64 Gaussian models. The confidence scores from different frames are summed to form the final classification decision.

III. Speech corpora

In this paper, we are mostly interested in distinguishing American accent from British accent. A normal practice is to use part of the same database for training and the other part for testing. One challenge is that the speech datasets are usually separately collected for the two accents. According to our knowledge, there is no database that collects both American English and British English speech data in the same effort. Thus, if we only use Database A for American English and Database B for British English, it is hard to say what the algorithm is used to distinguish the data from A and B. The distinction can be due to the intrinsic differences between the two accents, but it can also be the result of different external acoustic characteristics of the datasets, such as recording environment and equipment. Although Lincoln, et al, noticed this issue [2], they only tested their system on an independent American English database. In this paper, we use completely independent datasets for the testing and training of both British English and American English. Table I summarizes the datasets involved. We use WSJCAM0 for the training of British accent, IViE for the testing of British accent, TIDIGITS for the training of American accent, and Voicemail for the testing of American accent. In this way, we can truly examine the system's generalization capability on data collected in an environment completely different from the training data. One major reason to use Voicemail and IViE for testing is that the utterances in these two corpora have an average length of 32 and 42 seconds respectively. As shown in Section IV, the accent classifier's accuracy will saturate at such a length. Each utterance in the two corpora contains a short paragraph of several continuous sentences instead of a single sentence in many other frequently used speech corpora.

TABLE I

Speech Database Used in This Work

No	Name	Role in this paper	Description	Size
1	WSJCAM0 [9]	British accent training	UK English equivalent of a subset of the American English Wall Street Journal database	140 speakers
2	IViE [8]	British accent testing	British accent data collected by University of Cambridge for the study of intonational variation	45 speakers, totaling 36 hours of speech
3	TIDIGITS [10]	American accent training	Connected digits collected at Texas Instruments	326 speakers, each with about 77 utterances of digit strings
4	Voicemail [11]	American accent testing	Voicemail speech data collected by IBM	2200 voicemail messages, totaling 19.4 hours of speech

IV. Experimental Results And Analysis

We first compare the proposed hierarchical classification method with the direct classification without gender detection. The direct accent classification is similar to the phoneme-less accent classification except for a few differences: First, the GMMs are trained with blended male and female data. Second, 128 Gaussian models are used for American English and British English respectively. The number doubles that of the hierarchical scheme because the GMMs are expected to cover male and female simultaneously. From Table II, it can be seen that hierarchical classification reduces the error rate by 7.1% relatively compared with direct accent classification.

TABLE II

Error Rates On Voicemail And IViE Corpora

	Voicemail Male	Voicemail Female	IViE Male	IViE Female	Average
Hierarchical	22.3%	13.2%	16.1%	17.1%	17.2%
Direct	8.77%	20.5%	27.0%	17.7%	18.5%

The proposed hierarchical accent classification is closely related to gender detection, so it is interesting to compare the two problems. Intuitively, accent classification is much more difficult than gender detection for humans, so it should be the same for the computer. In fact, accents themselves are very complex linguistic and social phenomena [14], and the boundary between different accents can be very fuzzy. In order to verify this statement, we have compared the error rate of gender detection with that of accent classification on the same speech databases (see Fig. 3). As expected, the error rate of accent classification is much higher than that of gender detection rate.

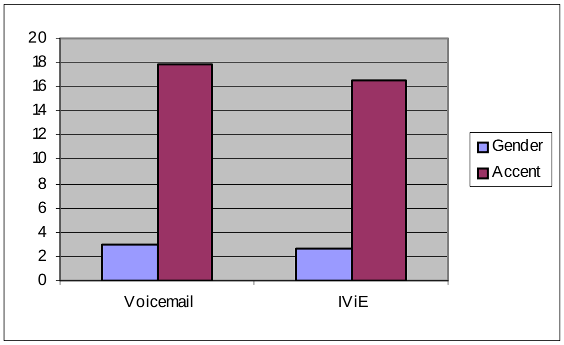

Fig. 3. Error rates of gender and accent classification

The above observations can be explained by the clustering characteristic of the samples in the feature space as illustrated in Fig. 4. Male and female samples are far way from each other, so they are easy to separate. In the male category, the samples can be subdivided into different accents. The same is true for the female category. Given this clustering characteristic, it is understandable why hierarchical classification performs better than direct classification.

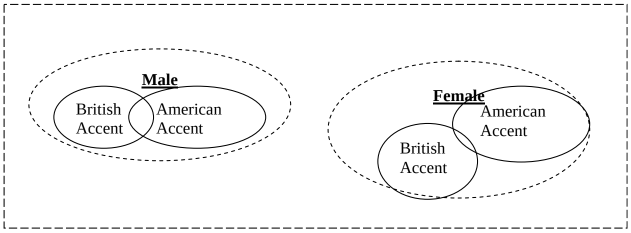

Fig. 4. Clustering characteristic

As shown in Fig. 2, the underlying accent recognition is based on the combination of the results from individual frames. Then there arises an interesting question: How does the length of an utterance affect the recognition rate? We have designed experiments to explore the asymptotic characteristic of accent recognition with the result displayed in Fig. 5. The X-axis corresponds to the length of utterances used. For example, only the first 5 seconds of each utterance is used in the first experiment. In the last experiment, the full length of each utterance is used. The average full length of utterances is 42 seconds for the IViE dataset and 32 seconds for the Voicemail dataset. The Y-axis is the error rate. The two curves are for IViE and Voicemail datasets respectively. As expected, for shorter utterance length, the more we use from each utterance, the lower the error rate is. However, the error rate starts to level off at around 30 seconds. In our previous work [6], it was the observed that the accuracy of gender detection levels off within the first second of utterance, which is much shorter than that of accent classification. This kind of difference actually exists in human listeners. We can usually decide a person's gender almost immediately while we have to listen to a couple of sentences in order to catch the characteristics of specific accents.

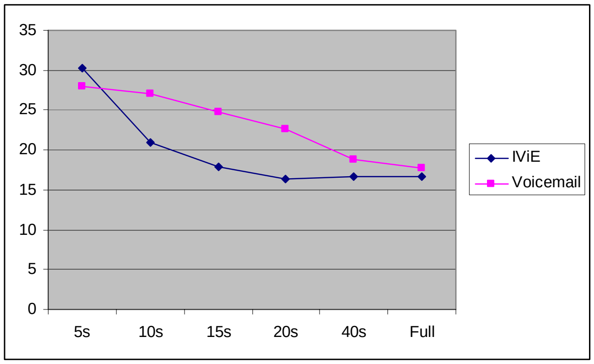

Fig. 5. Asymptotic characteristic of accent recognition

V. Application System

Based on the proposed algorithms, we have also built a VoiceSmart IVR application system. When the customers call in, the VoiceXML browser will invoke the accent and gender classifier. The classification result can then be used to customize sales offerings or advertisements. The server can visualize the classification results and statistics. Fig. 6 shows a screenshot.

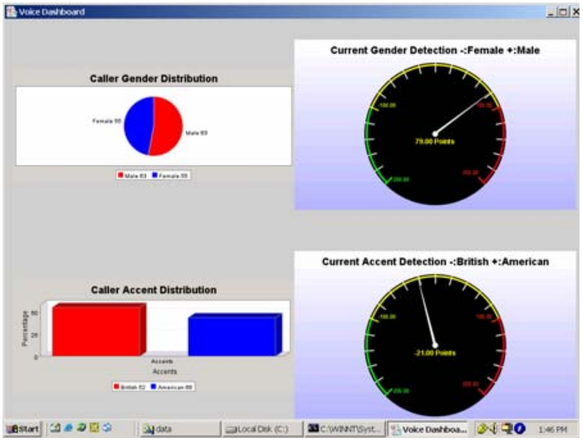
Voice Dashboard
Caller Gender Distribution
Female
Male
Current Gender Detection -:Female +:Male
79.00 Points
Caller Accent Distribution
Percentage
Accents
Current Accent Detection -:British +:American
-21.00 Points
Start
data
Local Disk (C:)
C:\WINNT\Syst...
Voice Dashboa...
1:46 PM

Fig. 6. Application system graphic user interface

VI. Conclusions

In this paper, we have introduced a new accent classification method. It is hierarchical and does not explicitly use phoneme information. Experimental results show that the hierarchical scheme outperforms the direct classification. More importantly, testing is conducted on large, independently collected speech databases to make sure that the results do reflect the real accent classification performance. We have also compared gender detection with accent classification and introduced the clustering relationship between the two tasks.

Because the proposed method does not use phoneme information, it is very suitable for efficient real-time implementation and it can be easily adapted to the classification of the accents of other languages. On the other hand, in order to "perfectly" distinguish accents, we need to look at level above pronunciation, such as lexical level (choice of words) and grammatical level (use of sentences) [14]. Of course, how to effectively (preferably automatically) extract such high-level knowledge and apply it to a feasible computational framework, especially when the raw input is speech signal instead of text, remains a challenge and one of our future research directions.

Acknowlegements

This research is conducted under the close collaboration with HP’s OpenCall Business Unit (OCBU). The authors are also grateful to Pedro Moreno and Beth Logan, who have provided us with the speech corpora.

References

[1] D. R. Miller and J. Trischitta, “Statistically Dialect Classification Based on Mean Phonetic Features,” *Proc. of ICSLP*, vol 4, Philadelphia, USA, 1996, pp. 2025-2027.

[2] M. Lincoln, S. Cox, and S. Ringland, “A Comparison of Two Unsupervised Approaches to Accent Identification,” *Proc. of ICSLP*, Sydney, Australia, December 1998.

[3] W. Labov, S. Ash, and C. Boberg, “A National Map of the Regional Dialects of American English,” http://www.ling.upenn.edu/phono_atlas /NationalMap/ NationalMap.html.

[4] P. Angkititrakul and J. H. L. Hansen, “Stochastic Trajectory Model Analysis for Accent Classification,” *Proc. of ICSLP, vol 1*, Denver, Colorado, USA, 2002, pp. 493-496.

[5] P. Fung and W. K. Liu, “Fast Accent Identification and Accented Speech Recognition,” *Proc. of ICASSP*, Phoenix, Arizona, USA, March 1999.

[6] X. Lin, “Decision Combination in Speech Metadata Extraction,” *Proc. of 37th Asilomar Conference on Signals, Systems and Computer*s, Pacific Grove, California, USA, November 2003.

[7] Carnegie Mellon University, “CMU Sphinx Speech Recognition System,” http://www.speech.cs.cmu.edu/sphinx.

[8] E. Grabe, B. Poat, and F. Nolan, “The IViE Corpus. Department of Linguistics,” University of Cambridge, http://www.phon.ox.ac.uk/ ~esther/ivyweb, 2001.

[9] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, “WSJCAM0 Corpus and Recording Description,” http://www.ldc. upenn.edu/Catalog/docs/LDC95S24/wsjcam0.html, 1995.

[10] R. G. Leonard and G. R. Doddington, "A Speaker-independent Connected-digit Database", http://www.ldc.upenn.edu/Catalog/docs /LDC93S10/tidigits.readme.html, 1993.

[11] M. Padmanabhan, B. Kingsbury, B. Ramabhadran, J. Huang, S. Chen, G. Saon, and L. Mangu, "Voicemail Corpus Part II", http://www.ldc. upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S35, 2002.

[12] M. D. Skowronski and J. G. Harris, “Increased MFCC Filter Bandwidth for Noise-robust Phoneme Recognition,” *Proc. ICASSP*, vol I, 2002, pp 801-804.

[13] X.D. Huang, A. Acero, and H.W. Hon, “*Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*,” Prentice Hall PTR. 2001.

[14] D. Crystal, “*The Cambridge Encyclopedia of the English Language,*” Cambridge University Press, 1995.