Automatic Accent Annotation with Limited Manually Labeled Data

Yining Chen[1] Min Lai[2] Min Chu[1] Frank K. Soong[1] Yong Zhao[1] Fangyu Hu[2]

[1]Microsoft Research Asia, Beijing, China
[2]Department of Electronic Engineering & Information Science, University of Science & Technology of China, Hefei, China
[1]{ynchen, minchu, frankkps, yzhao}@microsoft.com, [2]mlai@mail.ustc.edu, [2]hufy@ustc.edu.cn

Abstract

Annotating manually the accent labels of a large speech corpus is both tedious and time-consuming. In this paper we investigate automatic accent labeling procedure by using classifiers trained from limited manually labeled data. Different methods are proposed and compared in a framework of multi-classifiers, including: a linguistic classifier, an acoustic classifier and a combined one. The linguistic classifier is first used to label POS-determined content words as accented and function words as unaccented. The corresponding labels are then used to train accented and unaccented vowel HMMs separately. The combined classifier is then used to combine the decisions of the linguistic and acoustic classifiers' outputs to minimize labeling errors. Properly combined classifiers achieve better labeling performance than their linguistic and acoustic counterparts. The performance can be further improved when the acoustic classifier is re-trained with the whole corpus which is re-labeled by the combined classifiers. The final accent labeling accuracy is improved to 94.0%. Compared with 97.2%, the self-agreement ratio of a well-trained human annotator, this accuracy is fairly satisfactory.

1. Introduction

Labeling prosodic events in a speech database is important for both speech analysis and synthesis. Among all prosodic events, accent is probably the most prominent one. This paper focuses on how to detect and label accent automatically with classifiers trained on limited, manually labeled data. Several methods are compared to find the best utilization of limited manual labels. Before introducing the main content, we will review some necessary background first.

"What many phoneticians and linguists have called stress, and what most laymen readily understand under this term, refers to nothing more than the fact that in a succession of spoken syllables or words some will be perceived as more salient or prominent than others" [1]. Labeling accented syllables manually, especially for a very large speech database, is both labor-intensive and uneconomical. An efficient and reliable automatic prosody labeling scheme is highly desirable.

High intensity, long duration and high fundamental frequency are believed to be the primary acoustic cues for identifying accented syllables. Although how these three factors work together to make the accented syllables more prominent than the surrounding unaccented ones remains somewhat unclear, they have been commonly used to detect accents in previous studies [2, 3]. Accent is also found to be correlated with voice quality as well. Usually, accented vowels are pronounced more clearly than their unaccented counterparts who tend to be reduced. Hence spectral parameters such as Mel-scale Frequency Cepstral Coefficients (MFCC) are used in some accent detection studies [4]. Both [4] and [5] model the acoustic features of accented/unaccented vowels or syllables with Hidden Markov Models (HMMs).

When listening to an utterance, people not only use its acoustic but syntactic or semantic cues to help locating accents. Therefore, features derived from texts, such as part of speech (POS), N-Grams of POS and the positions within the phrase, are used in accent detections as well [5-7]. Bayesian decision [4] and artificial neural network (ANN) [6] have been employed to combine information at text and acoustic levels

The accuracy of accent prediction algorithms at word level is around 80-90% for different corpora and accent labeling methodologies. Most corpora used for the accent detection task are speaker independent [2-7]. As far as the labels used, ToBI (Tone and Break Index) are used in some cases and 3-4 levels of accent are labeled in others [5, 8].

In most of the studies, the classifiers used for marking accented/unaccented syllables are trained from the manually labeled data only. Due to the cost of labeling, the size of manually labeled data is often not large enough to train classifiers with high precision. How to improve the precision of a classifier with limited manual data is investigated in this paper. In a TTS speech corpus, there are often more unlabeled data available than the labeled ones. A possible way to improve the precision is to label these data automatically by employing a rough classifier (constructed with some prior, for example text level, information) and adding the automatically labeled data to the training set. Four different methods for combining unlabeled data with manual ones are introduced and their performances are compared in this paper.

In Section 2, a multiple classifier framework and its main modules are introduced. In Section 3, the four methods to use limited manually labeled data are described. Evaluations and results are presented in Section 4 and conclusions are outlined in Section 5.

2. Accent detection with multiple classifiers

In [9], we propose a multiple classifier framework for detecting accent. As illustrated in Fig. 1, the three classifiers are: an HMM-based acoustic classifier, a linguistic classifier and a combined one. The HMM-based acoustic classifier aims at exploiting the segmental information of accented vowels. The linguistic classifier aims at capturing the text level information. The combined classifier tries to bridge the mismatch between acoustic classifier and linguistic classifier with more accent related information like word N-gram scores, segmental duration and fundamental frequency differences among succeeding segments. The three classifiers are introduced below.

2.1. The linguistic classifier

According to Pike [1], usually content words which carry more semantic weight in a sentence are accented while function words are unaccented. Following this rule, a simple linguistic classifier is designed in this study: according to their POS tags, content words are deemed as accented while non-content or function words as unaccented.

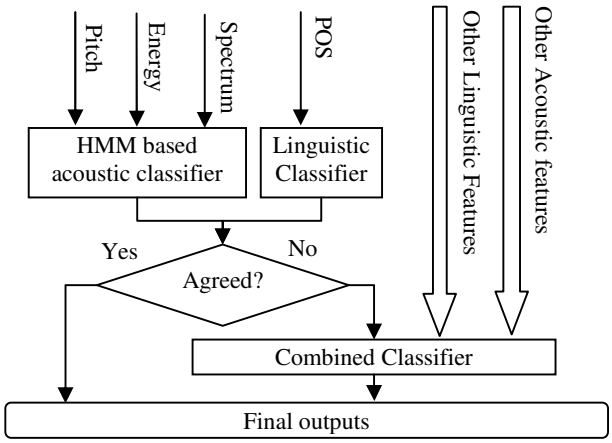

Figure 1: *The multiple classifier framework for accent detection*

2.2. HMM based acoustic classifier

The HMM based acoustic classifier uses the segmental information that can distinguish accented vowels from unaccented ones. First we need to choose a set of segmental units which are to be modeled.

2.2.1. Accent and position dependent phone set

In a conventional speech recognizer, about 40 phones are used in English and for each vowel a universal HMM is used to model both its accented and unaccented realizations. In our model the accented and unaccented are modeled separately as two different phones. Furthermore, to model the syllable structure which consists of onset, vowel nucleus and coda, with a higher precision, consonants at the onset position are treated differently from the same phones at the coda position. This accent and position dependent (APD) phone set increases the number of labels from 40 to 78 and but the corresponding HMMs can be trained similarly.

2.2.2. Training of APD HMMs

Before training the new HMMs, the pronunciation lexicon is adjusted in terms of the APD phone set. Each word pronunciation is encoded into both accented and unaccented versions. In the accented one, the vowel in the primary stressed syllable is accented and all the other vowels unaccented. In the unaccented word, all vowels are unaccented. All consonants at syllable-onset position are replaced with corresponding onset consonant models and similarly for consonants at coda position.

In order to train HMMs for the APD phones, accents in the training data have to be labeled, either manually or automatically. Details will be introduced in Section 3. Then, in the training process, the phonetic transcription of the accented version of a word is used if it is accented. Otherwise, the unaccented version is used.

Besides the above adjustment, the whole training process is the same as conventional speech recognition training. APD HMMs are trained with the standard Baum-Welch algorithm in the HTK software package [10]. The trained acoustic model is then used to label accents.

2.2.3. Accent labeling with APD HMMs

The accent labeling is actually a decoding in a finite state network as shown in Fig. 2 where multiple pronunciations are generated for each word in a given utterance. For monosyllabic words (as the 'from' in Fig. 2), the vowel has two nodes, A node (stands for the accented vowel) and U node (stands for the unaccented vowel). Each consonant has only one node, either O node (stand for an onset consonant) or C node (stand for a coda consonant). For multi-syllabic words, parallel paths are provided and each path has at most one A node (as in the word "city" in Fig. 2). After the maximum likelihood search, words aligned with accented vowel are labeled as accented and others as unaccented.

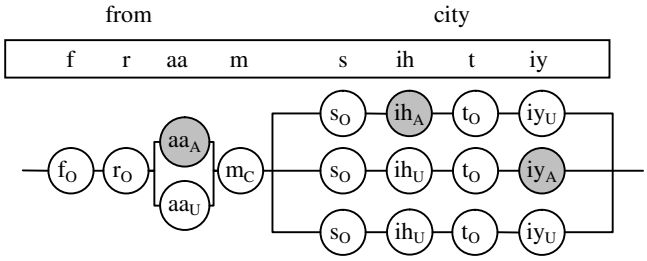

Figure 2: *Finite State Network for accent labeling.*

2.3. Combined classifier

Since the linguistic classifier and the acoustic classifier generate accent labels from different information sources, they do not always agree with each other. To reduce classification errors further, a third classifier can be constructed by combining the results of the above two via the AdaBoost algorithm with additional accent related, acoustic and linguistic information.

2.3.1. Features used by the combined classifier

Three accent related feature types are used. The first type consists of the likelihood scores of accented and unaccented vowel models and their differences. The second type addresses the prosodic features that cannot be directly modeled by the HMMs, such as the normalized vowel duration and fundamental frequency differences between the current and the neighboring vowels. The third type is the linguistic features beyond POS, like uni-gram, bi-gram and tri-gram scores of a given word because frequently used words tend to be produced with reduced pronunciations [11]. For each type of features, an individual classifier is trained first. However, its performance is not good. We then decide to combine these weak classifiers into a stronger one.

2.3.2. Combining Scheme

AdaBoost algorithm [12] is often used to adjust the decision boundaries of weak classifiers to minimize classification errors

and has resulted in better performance than each individual one [13]. The advantage of AdaBoost is that it can combine a sequence of weak classifiers by adjusting the weights of each classifier dynamically according to the errors in the previous learning step. In each boosting step, one additional classifier of single feature is incorporated.

3. Four ways to use the limited manual labels

When only a small number of manual labels are available, how to take the best advantage of them becomes crucial. Although it is not easy to train a high performance classifier with only limited manual labels, we can utilize the unlabeled data which are more abundant than their labeled counterparts to improve the labeling performance. In this section, several ways of using the manual labels and the unlabeled data are introduced. Under our accent detection framework, the manual labels are always used to train the combined classifier. The difference lies on the utilization of labeled or unlabeled data in training the acoustic classifier as shown in Table 1.

Table 1: *Data used in training acoustic classifier*

Data	Method 1	Method 2	Method 3
Manually Labeled data	√		√
Auto-labeled Data		√	√

As shown in the table, HMMs are trained only on the manually labeled data in method 1. The linguistic classifier described in Section 2.1 performs accent labeling of the whole speech corpus and the auto-labeled data are used in training HMMs in method 2. The accent labeling accuracy of the linguistic classifier is 91.6% which is good enough for training reasonable acoustic models. In method 3, both manual labels and auto labels are used in training HMMs. With the available manually labeled data, we hope the accuracy of the acoustic model to be better than that in method 2.

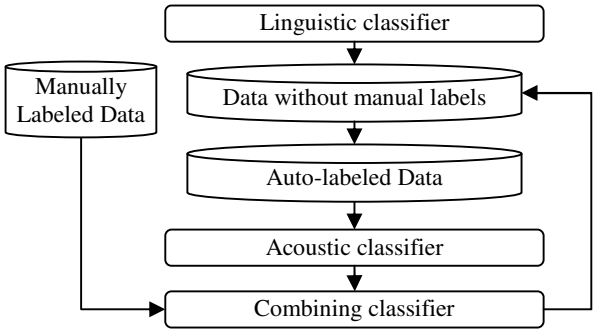

Figure 3: *Re-label the speech corpus with a more precise classifier.*

Although the linguistic classifier yields a 91.6% accent labeling accuracy on the test data, the combined classifier can do much better. Therefore, the fourth method shown in Fig. 3 is proposed. In method 4, the combined classifier is used to re-label the speech corpus and new acoustic models are further trained with the additional relabeled data.

4. Evaluation and results

In this section, experiments are performed to compare the accent labeling performance of four methods.

4.1. Experiment setup

The speech corpus we used contains 6,412 utterances, recorded by a professional female broadcaster. Accented words in the first 1,000 utterances have been labeled by a well trained annotator. These manual data are split evenly into training and testing sets. The 500-utterance training set is the "manually labeled data" and all other 5,412 sentences are the "auto-labeled data" in both Table 1 and Fig. 3.

The instruction given to the annotator is to label prominent words in the utterances by only listening to the utterances but not viewing the speech waveforms or spectrograms. The first 500 utterances are labeled twice by the same annotator with a 3 month time-span in between two attempts. The agreement ratio between the two labeling attempts is 97.2% which marks the upper bound for auto-labeling accuracy. In the following evaluation, accent labeling accuracy per word is measured.

4.2. Accuracy of the linguistic classifier

If all content words are labeled as accented and all function words as unaccented, the agreement ratio between linguistic classifier and human annotator is 91.6%, a fairly decent performance.

When we analyze the errors in content words, we found that most are high frequency words like "did" and "went." In many labeling errors of function words, multi-syllabic words like "every" and "around" are more likely to be labeled wrong.

4.3. Acoustic and Combined classifier Performance

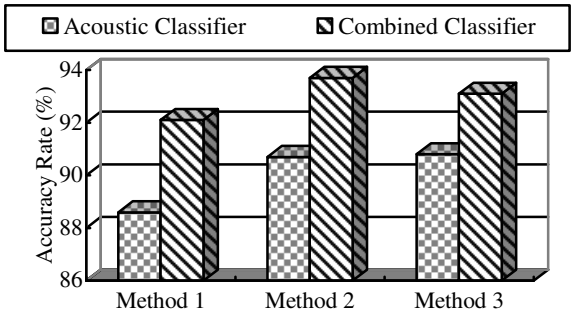

Figure 4: *Accuracy of acoustic classifier and combined classifier*

In Fig. 4, the labeling accuracies of acoustic classifiers and combined classifiers of the first three methods are shown. Labeling performance in terms of accuracy rate of the combined classifiers is significantly better than that of the acoustic classifiers. Methods 2 and 3 are better than method 1 for both classifiers. It indicates that when more (about 11 times) data is used, even with some labeling errors, better acoustic classifiers can be trained.

Although the accuracy of the acoustic classifier in method 3 is higher than that in method 2, the performance of the combined classifier in method 2 is better than that in method

3. This may due to the fact that when manual data is used to train a better acoustic classifier in method 3, the combined classifier can no longer benefit from the same manual data in a discriminative sense.

It is well known that a classifier can achieve better performance when the training scenario matches the testing one. In method 3, when the manually labeled dataset is used to train acoustic model, the discriminative training of the combined classifier can no longer simulate the testing scenario well. However, in method 2, since the manually labeled dataset is NOT part of the acoustic model training, the combined classifier is trained in a scenario closer to the real testing case than method 3. This may explain why method 2 performs better.

4.4. Accuracy of the four methods

Since the combined classifier in method 2 performs the best among the first three methods, we use it to re-label the whole corpus in method 4 shown in Fig. 3. The labeling accuracy of the combined classifiers for all methods is shown in Fig. 5. The method 4 outperforms all the other methods. It suggests that when training data is more accurate, an acoustic classifier with higher resolution can be trained. A 20% relative error reduction in the training data results in about 5% relative error reduction in testing data.

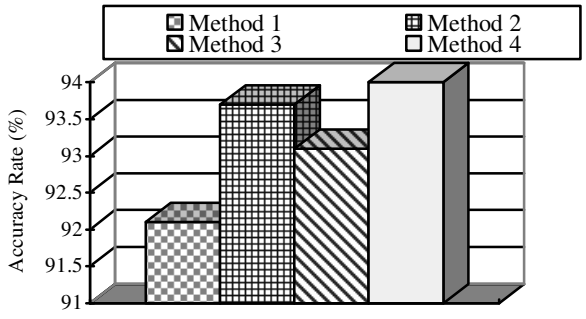

Figure 5: *Accuracy of method 1-4*

5. Conclusion

In this paper, several automatic accent annotation schemes are introduced to utilize limited manually labeled data and they are compared within a multiple-classifier framework. It is found that a better way to use the limited manual data is to keep it from training the acoustic classifier. An acoustic classifier with a high labeling precision can be trained with auto-labeled training data. Furthermore, increasing the precision of the rough labels by boosting can improve the performance of the acoustic classifier and the final results. The best accuracy we achieved is 94.0%, when using only 500 manually labeled sentences. This is fairly positive compared with the performance upper bound of 97.2%, which is the self agreement ratio of the human annotator. And it is much better than training classifiers with manual labels only (92.1%).

In this study, we are able to obtain good results with a model that labels all content as accented and all function words as unaccented. When tested on our TTS corpus, which is built up with isolated sentences, the result is quite acceptable. However, isolated sentences tend to be overly-accented since without contexts, say in a paragraph, lots of words carry "new" information which justifies their accentuation. In the future, we will continue with more complete experiments on longer speech segments like paragraphs.

6. Acknowledgments

The authors would like to thank Scott Meredith for his great help on creating the specification for prosody annotation. We would also like to offer special thanks to Yaya Peng for creating these accent labels.

7. References

[1] E. C. Kuhlen, "An Introduction to English Prosody", Edward Arnold, 1986.

[2] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," IEEE Trans. on Speech and Audio Processing, 2(4), pp. 469-481, 1994.

[3] I. Bulyko and M. Ostendorf. "A Bootstrapping Approach to Automating Prosodic Annotation for Constrained Domain Synthesis," in Proc. of the IEEE Workshop on Speech Synthesis, pp 115-118, 2002.

[4] A. Conkie, G. Riccardi, and R.C. Rose "Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events" in Proc. of EUROSPEECH, pp 523-526, 1999.

[5] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, "Modeling and automatic detection of English sentences accent for computer-assisted English prosody learning system", in Proc. of ICSLP, pp 749-752, 2002.

[6] K. Chen, and M. Hasegawa-Johnson, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in Proc. of ICASSP, pp 509-512, 2004.

[7] S. Arnfield, "Prosody and syntax in corpus based analysis of spoken English," Ph.D. dissertation, University of Leeds, Dec. 1994.

[8] P.C. Bagshaw. "Criteria for labelling prosodic aspects of English speech," In Proc. 4th. Australian International Conference on Speech Science and Technology, 1992.

[9] M. Lai, Y.N. Chen, M. Chu, etc, "A hierarchical approach to detect stress in English sentences", Submitted to ICASSP 2006.

[10] S. Young, G. Evermann, D. Kershaw, etc, "HTK Book, version 3.1", http://htk.eng.cam.ac.uk/prot-docs/htk_book.shtml

[11] S. Werner, etc, "Toward Spontaneous Speech Synthesis——Utilizing Language Model Information in TTS", IEEE Transactions on speech And Audio Processing, 12(4), pp 436-444, 2004.

[12] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," J. Comp. & Sys. Sci 55(1), pp 119-139, 1997.

[13] D. Wang, L. Lu, H.J. Zhang. "Speech Segmentation without Speech Recognition," in Proc. of ICASSP 2003, pp 468-471, 2003.