An Unsupervised Approach to Automatic Prosodic Annotation

Xinqiang Ni[*1], *Yining Chen*[2], *Frank K. Soong*[2], *Min Chu*[2], *Ping Zhang*[1]

[1]Institute of Electronics, Chinese Academy of Sciences, Beijing, China
[2]Microsoft Research Asia, Beijing, China

[1]xqni@mails.gucas.ac.cn, [2]{ynchen, frankkps, minchu}@microsoft.com, [1]pzhang@mail.ie.ac.cn

10.21437/Interspeech.2007-225

ABSTRACT

Accent is probably the most prominent part in prosodic events. Automatic accent labeling is important for both speech synthesis and automatic speech understanding. However, manually labeling data for traditional supervised learning is expensive and time consuming. In this paper, we propose an unsupervised learning algorithm to label accent automatically. First, we assume all content words are accented. We build an initial acoustic model with accented vowels in content words and high confidence unaccented vowels in function words. Then an iterative progress is executed to convergence. Experimental results show that this unsupervised learning algorithm achieves about 90% agreement on accent labeling. Compared with 84.3%, the accuracy of a typical linguistic classifier, a 30% relative error reduction is obtained.

Index Terms— **s**peech synthesis, prosodic annotation, unsupervised learning, function word, content word,

1. Introduction

Prosody labeling is important for both speech synthesis and automatic speech understanding. Among all prosodic events, accent is probably the most prominent. In this paper, because of the analysis of the acoustic behavior in function and content words, we introduce a novel method of auto-labeling without any manual labeled data.

There is extensive related work on accent auto-labeling. In [1], a syntax based decision-tree system is proposed, which achieves 82.4% speaker dependent accent labeling accuracy for Radio News. In [2], bi-gram models were used to predict accent from parts-of-speech (POS) with 91% accuracy. Meanwhile, several studies have been conducted on acoustic cues. High intensity, long duration and high fundamental frequency are found to be the primary acoustic cues for accented syllables. Therefore, they are used as the main acoustic features in the accent detection task in some studies [3, 4]. Accent is also found to correlate to voice quality as well. Hence, attempts using spectral parameters such as Mel-Scale Frequency Cepstral Coefficients (MFCC) are reported in [5]. In recent years statistical learning algorithms have been introduced for accent labeling. Bayesian decision [5] and artificial neural network (ANN) [6] has been employed to realize automatic prosody labeling. A three-layer hierarchical framework is proposed in [7]. Different methods using limited manually labeled data are proposed and compared in a framework of multi-classifiers in [8].

In most of these studies, classifiers used for labeling accented/unaccented syllables are trained from the manually labeled data. Due to labeling costs, the size of manually labeled data is hardly large enough to train classifiers with a high precision. Moreover, it is difficult to find the qualified people who can do required labeling work among customers who wish to build a voice for themselves. For the customs who want to build the voice for themselves, it is not easy to find the qualified people who can do this labeling work. Therefore, unsupervised methods such as k-means were studied recently [9].

In this paper, we study the relationship between POS and acoustic behavior of word accent. According to POS, words are classified to include function words and content words. Function words are the words with little inherent meaning but with important roles in the grammar of a language. Non-function words are called content words. Typically, they are nouns, verbs, adjectives, and adverbs. According to Pike [10], usually content words, which carry more semantic weight in a sentence, are accented while function words are unaccented. Following this rule, if we choose only to label all the content words as accented, accuracy will be as high as 96% in our corpus. While, for function words, the accuracy is only about 85%. Since the self-consistency of labels is about 97%, it is almost impossible to build a classifier that is better than this simple classifier. Function words are the only part we need to label and content words can be used as training set. We hope content words can be used in labeling function words. After the investigation, we find the accent vowels in content words and unaccented vowels in function words are suitable for constructing robust models. With these two models serving as the foundation of our study, we introduce an iteration method to improve performance. In our experiment of function word accent labeling, accuracy is 84.3%, if we label all words as unaccented words. Our new method increases accuracy to 89.0%.

In section 2, we introduced the Hidden Markov Model based (HMM based) acoustic classifier. Section 3 analyses the relationship between function words and content words. In section 4, we discuss the framework of our unsupervised learning algorithm in greater details. Our evaluations and results are presented in section 5 and conclusions are outlined in section 6.

[*] This work is done when the first author visits Microsoft Research Asia as an intern

2. HMM based acoustic classifier

2.1. Phone set

In a conventional speech recognizer, for each English vowel, a universal HMM is used to model both accented and unaccented realizations. In our system the accented (A) and unaccented (U) versions of the same vowel are trained separately as two different phones. For the consonant, there is only one version for each individual phoneme.

With the differences clear between content words and function words, accented and unaccented vowels are further split to accented (function word) AF, unaccented (function word) UF, accented (content word) AC and unaccented (content word) UC. There are 16 vowels, with each comparing four different versions. In total, the phone set includes 64 different vowels and 22 consonants. Tri-phone models are built based on this phone set.

2.2. HMM training

Linguistic studies suggest that all syllables but one in a word tend to be un-accented in continuously spoken sentences [11]. Hence, we constrain the maximum number of accented syllables to one per word. In an accented word, the vowel in the primary stressed syllable is accented and all the other vowels are unaccented. In an unaccented word, all vowels are unaccented.

Before HMM training, the pronunciation lexicon is adjusted in terms of the phone set. Each word pronunciation is encoded into both accented and unaccented versions. As Figure 1 shows, the phonetic transcription of the accented version of a word is used if it is accented. Otherwise, the unaccented version is used.

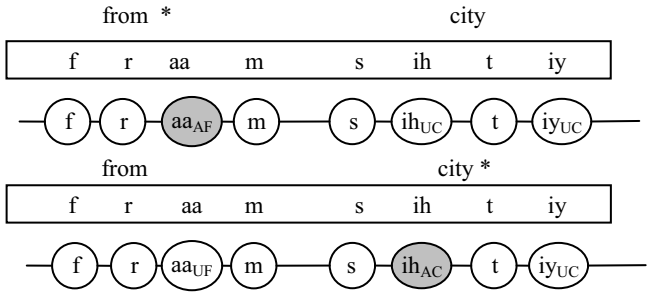

Figure 1: Accented and unaccented versions of pronunciation lexicon.

HMMs are trained with the standard Baum-Welch algorithm using the HTK software package [12]. The observations include MFCC and Pitch [7, 8]. The trained acoustic model is then used to label accent.

2.3. Accent labeling

The accent labeling is in fact a decoding process in a finite state network as shown in Figure 2. Multiple pronunciations are generated for each word in a given utterance. For monosyllabic words (as the word *from* in Figure 2), the vowel has two nodes, A node (stands for the accented vowel) and U node (stands for the unaccented vowel). For multi-syllabic words, parallel paths are provided; each path has at most one A node (as the word *city* in Figure 2). After maximum likelihood based decoding, words aligned with accented vowel are labeled as accented and others as unaccented. From this figure, we did not specify AF and AC from A nor UF and UC for U. We will discuss this problem in the next Section.

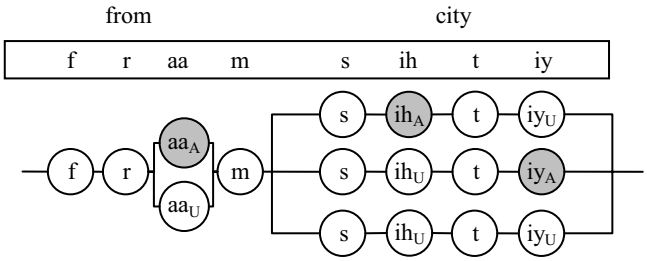

Figure 2: Finite state network for accent labeling.

3. Relationship between function and content words

In this section, we study the acoustic relationship between function and content words. That is the relationship between the function (AF UF) and content (AC UC) version of vowels. Each function word can be decoded by four models set with different combinations of four versions of vowels like Figure 3. Model (A): AC and UF. Model (B): AC and UC. Model (C): AC and UF. Model (D): AF and UC. All of these models are trained with supervised learning.

The accuracies of these four classifiers are shown in Figure 4. Supervised training is used to build the models. Obviously, the error rate of Model (A) is the best (12.0%), because function words are labeled by their own acoustic model. On the contrary, in Model (B), function words are labeled by acoustic models of content words. Thus the error rate increases to 38.0%. From the result of Model (C), we find the accent vowels in content words and unaccented vowels in function words can build a robust model. The error rate of function words is only 12.9%, almost identical to that of Model (A). The accuracy of Model (D) is non-optimal.

This observation is promising. As discussed in Section 1, we can use all the content words as a training set in our unsupervised approach. Hence, the AC and UC models can be found while Model B is the only model in the initial stage. If we can find enough unaccented vowels in function words, we can build a good UF model. Then with this UF model and the AC model got already, we can achieve the result similar to Model C, which is comparable to the best result of supervised learning.

In function words, about 85% are unaccented. It is not difficult to find enough unaccented vowels. When we further study the result of Model (B), although the accuracy of it is not good, the accuracy of unaccented labels is as high as 95.7%. That means most of the unaccented labeled in function words by Model (B) can be trusted as training set of model UF. Then the bridge between Model (B) and Model (C) is connected.

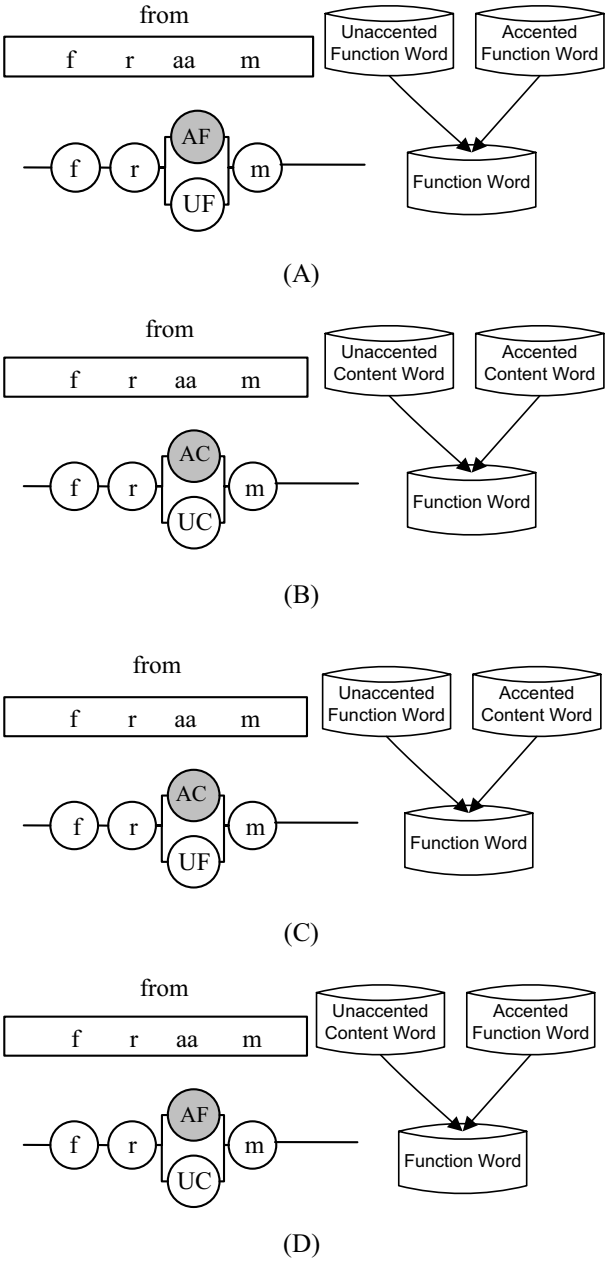

Figure 3: Accent labeling of four different methods.

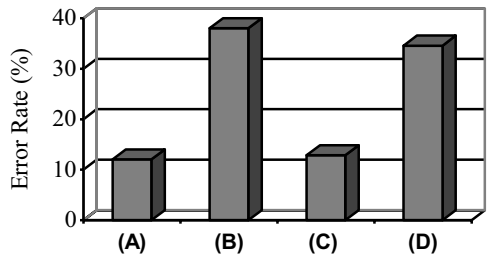

Figure 4: Performance of four models.

4. Unsupervised accent labeling

With the study in Section 3, we introduce an unsupervised training approach in Figure 5. Model (B') which is trained from all content words, is used to label the function words at the initial step. Vowels with the unaccented labels by Model (B') are used as the training set combined with all accented vowels in content words. Then, we train a Model (C') with this training set. It is quite similar to Model C. The only difference is it does not use all the unaccented vowels in function words. After the Model (C') is obtained, we can use it instead of Model (B') to repeat this work. When the labeling results of unaccented vowels in function words are very similar in two iterations, the training will stop. Finally, we achieve a converged Model (C').

Model B' and C' have the same structure with Model B and C in Figure 3. The difference is the models with prime are trained with unsupervised algorithm while the models without prime are trained with the supervised algorithm.

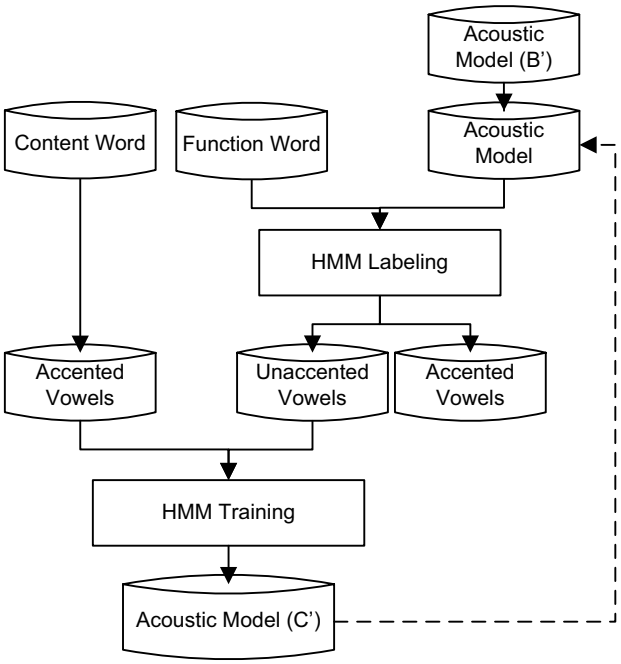

Figure 5: Unsupervised model training process.

5. Experiments and results

5.1. Experiment settings

The speech corpus used for evaluation consists of 6,412 utterances recorded by a professional female broadcaster. In the corpus, 3,000 utterances are labeled by a language expert. We use 1,000 labeled-sentences as testing data. In the unsupervised experiment, the 6,412 utterances without prosody labels are used as training set. In supervised experiment, 2,000 labeled-sentences are used as a training set.

As mentioned in Section 1, all content words can be viewed as accented words. Hence, in all the results, only those of function words are considered.

5.2. Experimental results

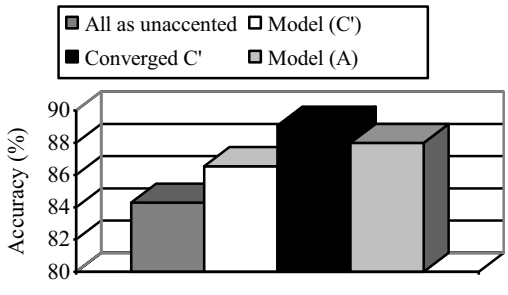

Figure 6: Results of different model.

Figure 6 shows the accuracy of our unsupervised learning algorithm. If all the function words are labeled as unaccented, the accuracy is 84.30%. When we use the Model (C'), the result is 86.6%, or a 14% relative error reduction. After the iteration, the accuracy of converged Model (C') increased to 89.1%, or a 30% relative error reduction.

When we compare this result to the supervised result of Model (A), we find this unsupervised algorithm is even better than the supervised result. Although we use less training data (2,000 sentences) in the supervised model, the result still show the ability of this algorithm.

6. Conclusions

In this paper we present an unsupervised learning algorithm for automatic sentence level accent labeling.

First, all the words can be classified into content and function words. We find that if we only label all the content words into accented words, the accuracy is close to the agreement ratio between labelers. Therefore, we can choose to focus on function words only.

Second, based on the relationship between function words and content words, we find using the accented vowels in content words and unaccented vowels in function words can build a good classifier. Since most function words are unaccented, the unaccented vowels in function words are easy to obtain.

Third, to find the unaccented vowels in function words, we use both the accented and unaccented vowels in content words to train a start up model. The unaccented output by this model to function words has a high level of accuracy. We can then get the unaccented vowels in function words we want. With these unaccented vowels and all the accented vowels in content words, a new model can be trained and 14% error reduction is obtained.

Last, the training progress detailed in the previous paragraph can be replicated. Each time, only the unaccented labels output by the classifiers are reused to train the new classifier. After the convergence, a 30% error reduction is achieved.

In the future, we will try to use additional features like durations and relative pitch to improve the results of this method.

7. Acknowledgments

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

8. References

[1] Hirschberg J., "Pitch Accent in Context: Predicting Intonational Prominence from Text," Artificial Intelligence, vol. 63 no. 1-2, 1993.

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

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

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

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

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

[7] Lai M., Chen Y.N., and Chu M., etc, "A Hierarchical Approach to Detect Stress in English Sentences", in Proc. of ICASSP, pp 753-756, 2006.

[8] Chen Y.N., Lai M., and Chu M., etc, "Automatic Accent Annotation with Limited Manually Labeled Data", in Proc. of Speech Prosody, 2006.

[9] Ananthakrishnan S., Narayanan S., "Combining Acoustic, Lexical, and Syntactic Evidence for Automatic Unsupervised Prosody Labeling", in Proc. of InterSpeech, pp 297-300, 2006.

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

[11] Watanabe K., "Instruction of English Rhythm and Intonation", chapter 4-6. Taishukanshoten, 1994.

[12] Young S., Evermann G., and Kershaw D., etc, "HTK Book, version 3.1", http://htk.eng.cam.ac.uk/protdocs/htk_book.shtml