Chinese Accent Detection Using Acoustic Feature Sets with Context Features

Zhao YunXue
College of Computer Science and Information Engineering
Harbin Normal University
Harbin, China
e-mail:zhaoyunxue_1126@163.com

Zheng ShiJie
College of Computer Science and Information Engineering
Harbin Normal University
Harbin, China
e-mail: zhengshijie@163.com

Zhang Long
College of Computer Science and Information Engineering
Harbin Normal University
Harbin, China
e-mail: zlwalkman@sina.com

Abstract—Accent is a critically important component of spoken communication. In this paper, we use acoustic features with context features to detect Chinese accent, and use NaiveBayes classifier to conduct model based on ASCCD reading discourse corpus. For comparison purposes, we digitally added white noise to ASCCD, and regarded it as ASCCD (white). NaiveBayes classifier is utilized to model the acoustic feature sets, which adequately present the property of the current syllable and context. The experimental results indicate that the acoustic feature sets with context features achieve 80.7% accent detection accuracy on ASCCD. The experiment also demonstrate that the acoustic feature sets with context features achieve 75.7% accent detection accuracy on ASCCD(white). Across these acoustic feature sets, we find pitch features to be the most predictive of Chinese accent on ASCCD and on ASCCD (white).Context is important to Chinese accent detection on ASCCD and on ASCCD (white).And the experimental results show that the NaiveBayes classification has a good classification result for these.

Keywords- Chinese accent detection; acoustic features; context features; NaiveBayes classifier;accent detection

I. INTRODUCTION

Accenting, or intonational prominence, has the effect of drawing a listener's attention to a particular section of a spoken utterance. Consider the following pair of utterances, where the words in quotation marks are accented.

My 'class' had a 'test' yesterday. The students performed 'horribly'.

The accenting of "class", "test" and "horribly" have the effect of highlighting the most salient information in these two utterances.

This accenting behavior is typically associated with an acoustic highlighting of a word, but the effect can be broader, drawing attention to the importance of a larger phrase. Take the following example, assumed to come after the two preceding utterances.

A student in the 'back' fell 'asleep' in the 'middle' of the exam.

The accenting of "back" may draw a listener's attention to the location of the student or indicate prominence of the whole noun phrase (NP), "A student in the back". The communicative effect of the accentuation of "back" is context dependent. If the previous utterance were discussing another student, the effect would likely be to draw contrast by highlighting the location of the subject of this utterance. On the other hand, without this context, accenting "back" serves to draw focus more broadly to the whole NP[1].

Fundamentally, accenting is an acoustic highlighting of a word through some modification of its associated speech signal. The effect of accenting, however, is not so simply understood. One theory posits that content words (e.g. nouns, verbs, adjectives, etc) introducing "new" information is accented, while terms conveying "given" or already established information are not accented, or deaccented[2-3]. This is something of an oversimplification, with empirical studies finding the existence of acoustically prominent "given" terms and non-prominent "new" terms[4-5]. Though not complete in its description of accenting behavior, this theory does have some explanatory power; "new" content words(nouns in particular) are often, if not always, accented, and "given" nouns are more likely not to bear accent.

One limitation of any theory of how accenting impacts the interpretation of an utterance, comes from the fact that accenting if used for multiple purposes. In addition to correlations with information status [2] contrast is used to indicate the focus[6] and topic[7] of an utterance. These three may suggest different accenting behavior for the same utterance. Consider the following utterance in our series of examples.

This 'class' is the 'worst' in the 'school'.

If information status were the only force impacting accenting behavior, "class" would not be accented-it was introduced to the discourse in the first utterance. However, the class is the topic and focus of the utterance. To make this clear to the listener "class" is accented. Accenting is also commonly used to highlight contrast, as in the following example.

'Your' class is a 'dream'!

"Your" is accented to draw a contrast between the speaker's and the listener's class. While the entity referred to by "your class" is new to the discourse, and the topic of the utterance, "Your" is accented, rather than "class" to highlight the contrast between the two classes.

There are also instances in which accenting more clearly affects meaning. Consider the following set of utterances.

1. The math department offers classes in geometry, algebra and trigonometry.

2A. Most sophomores have only 'heard' of algebra. (They haven't studied it.)

2B. Most sophomores have 'only' heard of 'algebra'. (They've never heard of geometry or trigonometry.)

The interpretation of utterances 2A and 2B significantly differ. Utterance 2A implies nothing about the student's relationship to geometry or trigonometry. On the other hand, 2B contrast their familiarity with algebra with a lack of awareness of the other material. While the complete breadth of communicative uses of accenting is not fully understood, there is consensus that accenting is an integral component of human speech.

This paper takes full advantage of features from acoustic (pitch, energy, intensity, and duration) and context, and constructs the Chinese accent detection feature sets. We use the weka machine learning method to train model, and then detect the Chinese accent.

This paper introduces research situation in the first part. The second part introduces ASCDD reading discourse corpus. The third part introduces the acoustic features with context features in detail. The fourth part describes experiment environment, and analyzes the experimental results. The experimental results and the development trend of this research field are given in the fifth part.

II. RELATED RESEARCH

Review of the domestic Chinese accent detection technology. Hu Weixiang et al[8] used the acoustic features of duration and pitch based on classification and regression tree structure model to detect Chinese accent. The accuracy can reach 78%. Shao Yanqiu et al[9]utilized neural network to analyze the acoustic features of Chinese accent detecting. The accuracy can reach 78.4%. NiChongJia et al[10] do the further research for the Chinese accent detecting. They made use of the acoustical relevant features and the grammar and dictionary relevant features to detect accent by Boosting integration classification and regression tree. And they used the grammar and dictionary relevant features to detect accent by condition random field constructing a model. And finally, they [11] integrated the Boosting classification and regression tree model and conditional random field model to gain higher recognition rate of hybrid model. It can achieve 76.3% on ASCCD corpus detection accuracy.

III. ASCCD

ASCCD applies to the study of language speech, speech engineering development and mandarin teaching. Corpus contains eighteen narrative texts and discussing discourses. Each text contains three to five paragraphs, and each paragraph has five hundred to six hundred syllables. A total of this corpus is nine thousand syllables. There have ten people, five women and five men. They are recorded as M001, M002, M003, M004, M005, F001, F002, F003, F004, F005. All syllables are annotated. Speech segment uses SAMPA - C standard labeling [12]. Rhythm uses the C - ToBI tagging system. It annotates pinyin, initial and final, tone and marking the syllable prosodic information level and sentence accent boundary. Accent of each unit of rhythmic is divided into level zero, one, two and three. Accent has rhythm structure corresponding Chinese hierarchy. The heaviest prosodic syllable weighs the phonetic symbol of one, the secondary prosodic phrase weighs the heaviest syllable the phonetic symbol of two, mainly the heaviest prosodic phrase syllable the phonetic symbol is three, zero indicates unstressed, namely normal pronunciation. In this study, we divide syllables into normal pronunciation and accent. We don't distinguish the differences among them. We consider prosodic word accent and the secondary prosody phrase (MIP) accent as a normal pronunciation, only the main rhythm of phrases (MAP) accent as stressed. ASCCD of accent distribution in the corpus as shown in table one.

TABLE I. DISTRIBUTION OF ACCENT IN ASCCD CORPUS

pronunciation	the number of syllables	the percentage
all	87586	100.00
normal pronunciation	76860	87.75
accent	10726	12.25

IV. ACCENT DETECTION USING ACOUSTIC FEATURE SETS

Literatures[13-14] indicate that duration, pitch and intensity have a strong correlation with accent. There is consensus and empirical confirmation that pitch, energy, intensity and duration is major acoustic correlates of Chinese accent. Therefore, this paper uses the duration, pitch, intensity, and energy related to predict the acoustic features of Chinese accent. We treat pitch accent detection as a binary classification problem, classifying words as accent bearing or non-bearing.

To establish baseline acoustic accent detection performance, we construct distinct feature sets using pitch, intensity, energy, duration, and voice quality features[15-18]. For the pitch and intensity feature set, we extract the minimum, maximum, mean, standard deviation, and z-score of the maximum pitch within the syllable from the raw and z-score speaker normalized pitch contours. The energy feature set extracts the same features from the raw and speaker normalized energy contours, as opposed to the pitch contour. The duration feature set consists of a single feature: in duration, in seconds, of the syllable.

Accents are regions of speech that are perceived to stand out from their surroundings. This prominence relative to the surrounding speech material suggests that acoustic features which capture contextual acoustic information should be able to improve Chinese accent detection[19-23]. To test this hypothesis, we extend each of the pitch, energy, and duration feature sets with context based features. For the pitch and energy features, the mean and maximum value within the syllable is z-score normalized using the mean and standard deviation from a number of context regions surrounding the current syllable. In the duration feature set, the value of the duration is z-

score normalized relative to the duration of the syllables comprising the context region. There are many monosyllabic words and two-syllable word in Chinese. For these normalizations, eight context-regions are defined: 1）one previous syllable,2）one following word,3）two previous word,4）two following word,5）one previous and one following word,6）two previous and one following word,7）one previous and two following words,8）two previous and two following words. This paper uses the algorithm of z-score for standardization characteristics.

V. TEST AND ANALYSIS OF EXPERIMENTAL RESULTS

A. *Experiment environments*

On ASCCD reading discourse corpus, we choose F001, F002, F003 and F005 as a training set, and select F004 as a test set. The size of the training set and testing set is 4:1 on sentence level. On the syllable level training set contains 35060 syllables, and the test set contains 8761 syllables, including about 964 accent syllables. For the machine learning method, we adopt the WEKA NaiveBayes classifier, and use the default settings for training.

Classification principle of NaiveBayes classifier is the prior probability through an object, using the NaiveBayes formula to calculate the probabilistic that the object belongs to. Then, it chooses classes with maximum a posteriori probability as the object's class.

For comparison purposes, we digitally added white noise to ASCCD, and regarded it as ASCCD (white). Other conditions are similar to experimenting on ASCCD. No noise and noise waveform diagram were shown below.

Figure 1. no noise speech signal waveform

Figure 2. white noise speech signal waveform

B. *Experimental results and analysis*

TABLE II. CLASSIFICATION ACCURACY USING ACOUSTIC FEATURES SETS

Feature Name	ASCCD	ASCCD(white)
Pitch	54.5%	54.5%
Intensity	30%	32.2%
Energy	41.2%	41.3%
Duration	29.8%	30.8%
All	63.8%	49.4%

Across these acoustic feature sets, we find pitch features to be the most predictive of Chinese accent on ASCCD and on ASCCD (white). These results support this of shenjiong[24-25].We also finds that intensity, energy and duration are weak predictors of Chinese accent. Pitch and energy features on ASCCD perform the same as on ASCCD (white). Intensity features on ASCCD perform better than on ASCCD (white).Duration features on ASCCD perform worse than on ASCCD (white).But the conclusion of intensity, energy and duration features improves the overall detection accuracy. The conclusion of pitch, intensity, energy and duration features reduce the overall detection accuracy on ASCCD (white).

TABLE III. CLASSIFICATION ACCURACY USING ACOUSTIC FEATURES SETS WITH CONTEXT FEATURES

Feature Name	ASCCD	ASCCD(white)
Pitch	76%	75%
Intensity	57.7%	59.2%
Energy	57.5%	50.6%
Duration	59%	41.2%
All	80.7%	75.7%

We find the inclusion of context normalized acoustic features to significantly improve automatic accent detect performance on ASCCD and on ASCCD (white). All four feature sets show improved performance with the addition of context features on ASCCD and on ASCCD (white) .We also find that pitch features to be the most predictive of Chinese accent on ASCCD and on ASCCD (white).The conclusion of pitch, intensity, energy and

duration features improve the overall detection accuracy on ASCCD and on ASCCD (white).The use of context in accent detection is well motivated by intonational theories-accented words is acoustically prominent. Prominence implies a differentiation from the norm; representing the surrounding acoustic material in the feature representations serves to capture this difference from the norm. The overall result is confirmed: context is important to Chinese accent detection on ASCCD and on ASCCD (white).

VI. CONCLUSION AND FUTURE WORK

On ASCCD reading discourse corpus, combining with the relevant acoustic characteristics of Chinese context multidimensional to detect accent, using the algorithm of NaiveBayes on current syllables and the around of acoustic characteristics to construct models, the method makes full use of the current syllables and the round of related properties. The experimental results show that the NaiveBayes classification has a good classification result. In the future, we want to simplify the characteristic and try to use a combination of other features, such as linguistic characteristics, and also to explore other modeling methods and techniques to depict the stress properties.

ACKNOWLEDGMENT

Our thanks to supports from the Natural Science Foundation of Heilongjiang Province of China (F201321) and the Application Technology Research And Development Project of Heilongjiang Province of China (GZ13A003). The authors are grateful for the anonymous reviewers who made constructive comments.

REFERENCES

[1] Hirschberg J. The pragmatics of intonational meaning [C].Speech Prosody 2002, International Conference. 2002.

[2] M.GriceandM.Savino.Can pitch accent type convey information statusin yes-no questions[C]. In Concept to Speech Generation Systems.1997.

[3] D.Dahan, M.Tanenhaus, andC.Chambers. Accent and reference solution in spoken-language Comprehension [J]. Journal of Memory and Language, 47:292-314, 2002.

[4] J.TerkenandJ.Hirschberg. Decantation of words representing 'given' information: Effects of Persistence of grammatical function and surface position [J].Language and Speech, 37(2):125-145, 1994.

[5] A.GravanoandJ. Hirschberg. Effect of genre, speaker and word class on the realization of given and New information. In Inter speech, pages 557-560, September 2006.

[6] J.Gundel. On different kinds of focus. InFocus: Linguistic, Cognitive and Computational Perspectives[C]. Cambridge University Press, 1999.

[7] N.Hedberg. The prosody of contrastive topic and focus in spoken English [C]. In Workshop on information Structure in context, 2003.

[8] Hu Weixiang, Dong Honghui, Taojianhua, Huang Taiyi.Study on stress perception in Chinese speech [J].Journal of Chinese Information Processing, 2005, 19(6):78-83.

[9] Shao Yanqiu, Han Jiqing, Liu ting, Zhao Yongzhen.Study on automatic prediction of sentential stress with natural style in Chinese [J].Acta Acustica, 2006, 31 (3):203-210.

[10] Ni Chongjia, Zhang Aiying, Liu Wenju. Mandarin Stress Detection Using Acoustic, Lexical and Syntactic Features [J].Chinese Journal of Computers, 2011, 34(9)1638-1647.

[11] Ni Chongjia,Liu Wenju,Xu Bo. Mandarin Stress Detection based complementary model[J].Computer Enginnering,2011,37(23)20-23.

[12] Chen Xiao-Xia, Li Ai-Jun, Sun Guo–Hua, Wu Hua, Yin Zhi-Gang. An application of SAMPA-C for standard Chinese[C]. Proceedings of the International Conference on Spoken Language Processing. Beijing, China, 2000:652-655.

[13] Pitrelli J F. ToBI prosodic analysis of a professional speaker of American English [J].Proceedings of the Speech Prosody,Nara, Japan, 2004:557-560.

[14] Nenkova A, Brenier J, Kothari A et al. To memorize or to Predict: Prominence labeling in conversational speech [J].Proceedings of the HLT-NAACL,Rochester, NY, USA, 2007:9-16.

[15] Kim D K, Chang J H. Statistical voice activity detection in kernel space[J]. The Journal of the Acoustical Society of America, 2012, 132(4):303–309.

[16] Chang J H. Statistical Model-Based Voice Activity Detection Based on Second-Order Conditional MAP with Soft Decision[J]. ETRI Journal, 2012, 34(2):184–189.

[17] Deng S, Han J, Zheng T, et al. A modified MAP criterion based on hidden Markov model for voice activity detecion[C]//IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). Prague: IEEE, 2011:5220–5223.

[18] Omar M K. Speech Activity Detection for Noisy Data Using Adaptation Tech-niques[C]//Proceedings of Interspeech. Portland, Oregon: Curran Associates, Inc.,2012:1373–1376.

[19] Ying D, Yan Y, Dang J, et al. Voice activity detection based on an unsupervised learning framework[J]. IEEE Transactions on Audio, Speech, and Language Pro-cessing, 2011, 19(8):2624–2633.

[20] Cover T M, Thomas J A. Elements of information theory[M]. Canada: Wiley-interscience, 2012.

[21] Ghosh P K, Tsiartas A, Narayanan S. Robust voice activity detection using long-term signal variability[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(3):600–613.

[22] Goodwin G C, Sin K S. Adaptive filtering prediction and control[M]. Courier:Dover Publications, 2013.

[23] Donoho D L, Elad M, Temlyakov V N. Stable recovery of sparse overcomplete rep-resentations in the presence of noise[J]. IEEE Transactions on Information Theory,2006, 52(1):6–18.

[24] Shen jiong, Hoek J H. Chinese language potential accents voice management (summary report)[J]. Language Studies,1994, (3):10-15.

[25] Motter B C. Focal attention produces spatially selective processing in visual cor-tical areas V1, V2, and V4 in the presence of competing stimuli[J]. Journal of Neurophysiology, 1993, 70(3):909–919.