Classification of Prosodic Events using Quantized Contour Modeling

Andrew Rosenberg
Department of Computer Science
Queens College CUNY, New York, USA
andrew@cs.qc.cuny.edu

Abstract

We present Quantized Contour Modeling (QCM), a Bayesian approach to the classification of acoustic contours. We evaluate the performance of this technique in the classification of prosodic events. We find that, on BURNC, this technique can successfully classify pitch accents with 63.99% accuracy (.4481 CER), and phrase ending tones with 72.91% accuracy.

1 Introduction

Intonation can significantly vary the intended meaning of a spoken utterance. In Standard American English, contrast is frequently indicated with an accent that has a steeper pitch rise – "I went to the **store** (not the library)" – than an accent that is used to indicate focus or introduce new information – "I went to the **store** (before going home)" . At phrase boundaries, rising pitch can indicate uncertainty or that the speaker is asking a question – "John likes Mary?" vs. "John likes Mary". Automatically detecting prosodic events and classifying their type allows natural language understanding systems access to intonational information that would unavailable if processing transcribed text alone.

The ToBI standard of intonation (Silverman et al., 1992) describes intonational contours as a sequence of High and Low tones associated with two types of prosodic events – pitch accents and phrase boundaries. The tones describe an inventory of *types* of prosodic events. In this work, we present Quantized Contour Modeling, a novel approach to the automatic classification of prosodic event types.

In Section 2, we describe related work on this task. We describe Quantized Contour Modeling in Section 3. Our materials are described in Section 4. Experimental results are presented and discussed in Section 5. We conclude and describe future directions for this work in Section 6.

2 Related Work

Five types of pitch accents – pitch movements that correspond to perceived prominence of an associated word – are defined in the ToBI standard(Silverman et al., 1992): H*, L*, L+H*, L*+H, H+!H*. In addition to these five, high tones (H) can be produced in a compressed pitch range indicated by (!H). For the purposes of the experiments described in this paper, we collapse high (H) and downstepped High (!H) tones into a single class leaving five accent types. The ToBI standard describes two levels of phrasing, intermediate phrases and intonational phrases which are comprised of one or more intermediate phrases. Each intermediate phrase has an associated phrase accent describing the pitch movement between the ultimate pitch accent and the phrase boundary. Phrase accents can have High (H-), downstepped High (!H-) or low (L-) tones. Intonational phrase boundaries have an additional boundary tone, to describe a final pitch movement. These can be high (H%) or low (L%). Intonational phrases have five possible phrase ending tone combinations, L-L%, L-H%, H-L%, !H-L% and H-H%. In section 5.3, we describe experiments classifying these phrase ending tones.

The *detection* of pitch accents and phrase boundaries has received significantly more research attention than the *classification* of accent types and phrase ending behavior. However, one technique that has been used in a number of research efforts is to simultaneously detect and classify pitch accent. This is done by representing pitch accent detection and classfication as a four-way classification task, where a token may be classified as UNACCENTED, HIGH, LOW, or DOWNSTEPPED. Both Ross and Ostendorf (1996) and Sun (2002) used this approach, reporting 72.4% and 77.0% accuracy respectively when evaluated on a single speakers. Levow also used this four-way classification for pitch accent detection and classification under supervised (2005), and unsupervised and semi-supervised learning approaches (2006). Using

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 721–724,
Los Angeles, California, June 2010.

SVMs with only acoustic features, 81.3% accuracy at the syllable level is achieved. Using unsupervised spectral clustering, 78.4% accuracy is reported, while using the semi-supervised technique, Laplacian SVMs, 81.5% accuracy is achieved. Since these approaches simultaneously evaluate the detection *and* classification of pitch accents, direct comparison with this work is impossible.

Ananthakrishnan and Narayanan (2008) used RFC (Taylor, 1994) and Tilt (Taylor, 2000) parameters along with word and part of speech language modeling to classify pitch accents as H*, !H*, L+H* or L*. When evaluated on six BURNC speakers using leave-one-speaker-out cross-validation, accuracy of 56.4% was obtained. In the same work, the authors were able to classify L-L% from L-H% phrase-final tones in the BURNC with 67.7% accuracy. This performance was obtained using RFC F0 parameterization and a language model trained over categorical prosodic events.

3 Quantized Contour Modeling

In this section, we present a modeling technique, Quantized Contour Modeling. This technique quantizes the f0 contour of a word in the time and pitch domains, generating a low-dimensional representation of the contour. The pitch of the contour is linearly normalized to the range between the minimum and maximum pitch in the contour, and quantized into N equally sized bins. The time domain is normalized to the range [0,1] and quantized into M equally sized bins. An example of such a quantization is presented in Figure 1 where $N = 3$ and $M = 4$. Using this quantized representation of a pitch contour, we

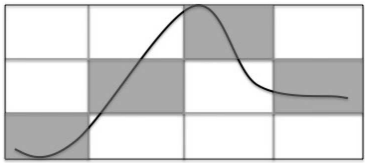

Figure 1: *Quantization with N=3 value and M=4 time bins.*

train a multinomial mixture model for each pitch accent type. Let the quantized contour be an M dimensional vector C where $C = (C_1, C_2, \ldots, C_M)$, where $C_i \in \{0 \ldots N-1\}$. We indicate pitch (f0) contours by C^{f0} and intensity contours by C^I. We train a multinomial model $p(type|C_i, i)$ for each time bin $i \in \{0 \ldots N-1\}$ with Laplace (add-one) smoothing. When using multinomial models, we quantize the mean of the pitch values assigned to a time bin. We use these pitch accent type models to classify a contour using the Bayesian classification function found in Equation 1. This formulation assumes that the values at each time are conditionally independent given the contour type. Also, we can modify the model incorporating a Markov hypothesis to include a sequential component by explicitly modeling the current and previous quantized values, as in Equation 2. We extend each of these models to model the energy contour shape simultaneously with the pitch contour. The classification technique allows for the number of pitch and energy value quantization bins to be distinct. However, in these experiments, we tie these, constraining them to be equal. The form of the classification functions using the energy contours are found in Figure 2.

Standard shape modeling

$$type^* = \underset{type}{\operatorname{argmax}}\, p(type) \prod_i^M p(C_i|type, i) \quad (1)$$

Sequential f0 modeling

$$type^* = \underset{type}{\operatorname{argmax}}\, p(type) \prod_i^M p(C_i|C_{i-1}, type, i) \quad (2)$$

Standard f0 + I modeling

$$type^* = \underset{type}{\operatorname{argmax}}\, p(type) \prod_i^M p(C_i^{f0}, C_i^I|type, i) \quad (3)$$

Sequential f0 + I modeling

$$type^* = \underset{type}{\operatorname{argmax}}\, p(type) \prod_i^M p(C_i^{f0}, C_i^I|C_{i-1}^{f0}, C_i^I, type, i) \quad (4)$$

Figure 2: *Quantized contour modeling classification formulae.*

4 Materials and Methods

We use two corpora that have been manually annotated with ToBI labels to evaluate the use of QCM in the classification of prosodic events. These two corpora are the Boston University Radio News Corpus (BURNC) (Ostendorf et al., 1995) and the Boston Directions Corpus (BDC) (Nakatani et al., 1995). The BURNC is a corpus of professionally read radio news data. A 2.35 hour, 29,578 word, subset from six speakers (three female and three male) has been prosodically annotated. The BDC is made up of elicited monologues spoken by four non-professional speakers, three male and one female. The BDC is divided into two subcorpora comprised of spontaneous and read speech. The 50 minutes of read speech contain 10,831 words. There are 60 minutes of annotated spontaneous material containing 11,627 words. Both are spoken by the same four speakers. In these experiments we evaluate these subcorpora separately, and refer to them as BDC-spon and BDC-read, respectively. The distribution of pitch accents and phrase-ending tones for these three corpora can be found in Figure 3.

Corpus	H*	L+H*	L*	L*+H	H+!H*
BDC-read	78.24%	13.72%	5.97%	1.36%	0.71%
BDC-spon	84.57%	6.32%	7.70%	0.68%	0.73%
BURNC	69.99%	21.64%	3.67%	0.34%	4.37%

Corpus	L-L%	L-H%	H-L%	!H-L%	H-H%
BDC-read	49.00%	35.62%	9.66%	4.29%	1.43%
BDC-spon	29.45%	32.57%	30.96%	4.40%	2.61%
BURNC	56.16%	38.38%	3.57%	0.68%	1.20%

Figure 3: *Distribution of prosodic event types in BURNC, BDC-read and BDC-spon corpora.*

In order to use QCM classification, we must first identify the region of an acoustic contour to quantify. Though there is evidence that acoustic evidence of prominence crosses the syllable boundary (Rosenberg and Hirschberg, 2009), it is largely held that the acoustic excursion corresponding to intonational prominence is centered around a syllable. To identify the region of analysis for QCM, we identify the accent-bearing syllable from the manual prosodic annotation, and quantize the contour extracted from the syllable boundaries. For the BURNC material, forced alignment syllable boundaries are available. However, no forced-alignment phone information is available for the BDC data. Therefore we apply Villing et al.'s (2004) envelope based pseudosyllabification routine to identify candidate syllabic regions. We use the pseudosyllable containing the accent annotation as the region of analysis for the BDC material. For classification of phrase ending intonation, we use the final syllable (or pseudosyllable) in the phrase as the region of analysis. To be clear, the accent and phrase boundary locations are derived from manual annotations; the intonational tones associated with these events are classified using QCM.

5 Prosodic Event Classification Results

In this section we present results applying QCM to the classification of pitch accents and phrase ending intonation. The work described in this section assumes the *presence* of prosodic events is known *a priori*. The approaches described can be seen as operating on output of an automatic prosodic event *detection* system.

5.1 Combined Error Rate

Automatic pitch accent classification poses an interesting problem. Pitrelli, et al. (Pitrelli et al., 1994) report human agreement of only 64.1% on accent classification in the ToBI framework. If downstepped variants of accents are collapsed with their non-downstapped forms this agreement improves to 76.1%. Second, pitch accents are overwhelmingly H* in most labeled corpus, including the BDC and BURNC material used in this paper. This skewed class distribution leads to a very high baseline, at or above the rate of human agreement. Because of this, we find accuracy an unreliable measure for evaluating the performance of this task. Multiple solutions can have similar accuracy, but radically different classification performance on minority classes. We therefore propose to use a different measure for the evaluation of pitch accent type classification. We define the Combined Error Rate (CER) as the mean of the weighted rates of Type I and Type II errors. The combination of these measures results in an increased penalty for errors of the majority class while being more sensitive to minority class performance than accuracy. Throughout this chapter, we will continue to report accuracy for comparison to other work, but consider CER to provide a more informative evaluation. To avoid confusion, accuracy will be reported as a percentage (%) while CER will be reported as a decimal.

$$CER = \frac{p(FP) + p(FN)}{2} \tag{5}$$

The Type I error rate measures the false positive rate for a given class (cf. Equation 6).

$$p(FP) = \sum_i p(C_i)p(FP_i) \tag{6}$$

We combine this measure with the Weighted Type II Error Rate (cf. Equation 7). The Type II error rate measures the false negative rate for a given class

$$p(FN) = \sum_i p(C_i)p(FN_i) \tag{7}$$

5.2 Pitch Accent Classification

The first step in applying Quantized Contour Modeling is to fix the desired quantization parameters. We do this by identifying a stratified 10% held out tuning set from the training data. We evaluate quantization sizes ranging between 2 and 7 for both the time and value parameters, leading to 36 candidates. Once we identify the best parameterization on this tuning data, we run ten-fold cross validation on the remaining data to evaluate the performance of each modeling technique (cf. Figure 2).

The classification accuracy and CER for each model is reported in Table 1 along with the number of time and value bins that were used. We first observe that modeling intensity information with f0 data does not improve classification performance. The alignment between pitch and intensity peaks have been shown to distinguish pitch accent types (Rosenberg, 2009); this relationship is not successfully captured by QCM. Moreover, we find that sequential modeling only leads to improvements in CER on BDC-read. On all corpora, the classification accuracy is improved, with statistically insignificant ($p > 0.05$) reductions in CER. This leads us to consider sequential modeling of pitch to be the best performing approach to the classification of pitch accent using QCM.

Method	BDC-read	BDC-spon	BURNC
f0	46.51/.3860(5,3)	**55.41/.4103(3,4)**	**47.56/.4444(4,4)**
Seq. f0	**73.17/.3667(6,7)**	81.20/.4156 (7,5)	63.99/.4481(7,7)
f0+I	37.53/.4094(3,3)	47.96/.4222(4,2)	48.36/.4472(2,2)
Seq. f0+I	74.08/.4032(7,3)	80.60/.4361(5,4)	66.97/.4530(6,5)
Baseline	78.22/.0000	84.57/.0000	70.23/.0000

Table 1: *Accuracy (%), CER, time and value bins from QCM pitch accent type classification experiments.*

5.3 Phrase-ending Tone Classification

As in Section 5.2, we identify the best performing quantization parameters on a stratified 10% tuning set, then run 10-fold cross validation on the remaining data. Results from QCM classification experiments classifying intonational phrase ending tone combinations – phrase accent and boundary tone – can be found in Table 2. We find

Method	BDC-read	BDC-spon	BURNC
f0	48.21(3,6)	40.26(2,2)	70.36 (5,2)
Seq. f0	53.86(2,2)	43.80(4,4)	71.77 (6,2)
f0+I	48.21(6,6)	38.28(6,6)	67.83(2,2)
Seq. f0+I	**57.94(6,6)**	**46.61(6,5)**	**72.91(7,7)**
Baseline	49%	32%	55%

Table 2: *Accuracy (%), time and value bins from QCM phrase ending tone classification experiments.*

that the simultaneous modeling of f0 and intensity consistently yields the best performance in the classification of phrase ending tones. These results all represent significant improvement over the majority class baseline. The interaction between pitch and intensity contours in the classification of phrase-ending intonation has not been thoroughly investigated and remains an open area for future research.

6 Conclusion and Future Work

In this paper we present a novel technique for the classification of two dimensional contour data, Quantized Contour Modeling (QCM). QCM operates by quantizing acoustic data into a pre-determined, fixed number of time and value bins. From this quantized data, a model of the value information is constructed for each time bin. The likelihood of new data fitting these models is then performed using a Bayesian inference.

We have applied QCM to the tasks of classifying pitch accent types, and phrase-ending intonation. The best performing parameterizations of QCM are able to classify pitch accent types on BURNC with 63.99% accuracy and $.4481$ Combined Error Rate (CER). QCM classifies phrase ending tones on this corpus with 72.91% accuracy.

These results do not represent the best performing approaches to these tasks. The best reported classification of pitch accent types on BURNC is 59.95% accuracy and .422 CER, for phrase ending intonation 75.09% (Rosenberg, 2009). However, the classification of phrase ending intonation is accomplished by including QCM posteriors in an SVM feature vector with other acoustic features.

This technique may be applicable to classifying other phenomena. Here we have used ToBI tone classifications as an intermediate representation of intonational phenomena. QCM could be used to directly classify turn-taking behavior, or dialog acts. Also, previous work has looked at using the same techniques to classify prosodic events and lexical tones in tonal languages such as Mandarin Chinese. QCM could be directly applied to lexical tone modeling; the only modification required would be a different segmentation routine.

References

S. Ananthakrishnan and S. Narayanan. 2008. Fine-grained pitch accent and boundary tone labeling with parametric f0 features. In *ICASSP*.

G.-A. Levow. 2005. Context in multi-lingual tone and pitch accent recognition. In *Interspeech*.

G.-A. Levow. 2006. Unsupervised and semi-supervised learning of tone and pitch accent. In *HLT-NAACL*.

C. Nakatani, J. Hirschberg, and B. Grosz. 1995. Discourse structure in spoken language: Studies on speech corpora. In *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*.

M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. 1995. The boston university radio news corpus. Technical Report ECS-95-001, Boston University, March.

J. Pitrelli, M. Beckman, and J. Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the tobi framework. In *ICSLP*.

A. Rosenberg and J. Hirschberg. 2009. Detecting pitch accents at the word, syllable and vowel level. In *HLT-NAACL*.

A. Rosenberg. 2009. *Automatic Detection and Classification of Prosodic Events*. Ph.D. thesis, Columbia University.

K. Ross and M. Ostendorf. 1996. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech & Language*, 10(3):155–185.

K. Silverman, et al. 1992. Tobi: A standard for labeling english prosody. In *ICSLP*.

X. Sun. 2002. Pitch accent predicting using ensemble machine learning. In *ICSLP*.

P. Taylor. 1994. The rise/fall/connection model of intonation. *Speech Commun.*, 15(1-2):169–186.

P. Taylor. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*.

R. Villing, et al. 2004. Automatic blind syllable segmentation for continuous speech. In *ISSC*.