TONE AND PITCH ACCENT CLASSIFICATION USING AUDITORY ATTENTION CUES

Ozlem Kalinli

US R&D, Sony Computer Entertainment, Foster City, California, USA.
e-mail: ozlem_kalinli@playstation.sony.com

ABSTRACT

A detailed description of tone and intonation is beneficial for many spoken language processing applications. In traditional methods for tone and pitch accent modeling, prosodic features, such as pitch, energy and duration, have been used. Here, a novel system that uses auditory attention cues is proposed for tone and fine grained pitch accent classification. The auditory attention cues are biologically inspired and hence extracted by mimicking the processing stages in the human auditory system. When tested on the Boston University Radio News Corpus, the proposed method achieves 64.6% pitch accent and 89.7% boundary tone classification accuracy. In addition, it is demonstrated that the model also successfully recognizes lexical tones in Mandarin with 79.0% accuracy when tested on a continuous Mandarin Chinese speech database. The results compare very well to the reported human performance on these tasks.

Index Terms— pitch accent, boundary tone, lexical tone, tone recognition, auditory attention, auditory gist.

1. INTRODUCTION

Tone and intonation play an important role in speech recognition and natural language understanding. In many languages, intonation, variation of pitch while speaking, can be used for emphasis, posing a question, or conveying surprise, etc. Hence, intonation can alter the meaning of a spoken utterance. For example, in standard American English, rising pitch at the end of a phrase often indicates that the speaker is asking a question instead of a declaring a statement ("He bought a car?" vs. "He bought a car"). Different from English and some western languages, tone languages such as Chinese use pitch to distinguish words. In tone languages, syllables or words which have the exact same sequence of phonemes often map to different lexical entries when they have different tone patterns or in other words pitch contours. For example, the words "mama"(mā), "hemp"(má), "horse"(mǎ), and "curse"(mà) are all pronounced "ma" in Mandarin, but each of them has a different tone pattern. Because of aforementioned reasons, a detailed description of tone and intonation would be beneficial for many spoken language processing systems; i.e. to disambiguate words in automatic speech recognition, to detect different speech acts in dialog systems, to generate more naturally sounding speech in speech synthesis systems, etc. Hence, here, we focus on recognition of lexical tones in Mandarin and fine-grained intonation types, namely pitch accent and boundary tones, in English.

In the literature, for English, the detection of pitch accent and boundary tones has been largely explored; however the classification of pitch accent and boundary tone types hasn't been investigated much until recent years. [1, 2, 3] combined pitch accent detection and classification tasks by creating a four-way classification problem with unaccented, high, low, and downstepped accent categories. Recently, [4] and [5] focused solely on the classification of pitch accent and boundary tone categories without the worry of detection. In contrast to fine-grained intonation classification in English, lexical tone recognition in Mandarin has attracted attention of researchers for many years, and approaches can be grouped under two major categories; namely embedded and explicit tone modeling. In embedded tone modeling, tone related features are augmented to spectral features at each frame and tone is recognized as a part of the existing system [6], whereas in explicit tone modeling, tones are independently modeled and recognized usually using supra-segmental features [7, 3, 8].

In traditional methods, including aforementioned work, prosodic features, which consist of pitch, duration and energy, were used for tone and intonation modeling. Here, a novel method that uses auditory attention cues for tone and pitch accent classification is proposed. The model shares the same front-end processing with the bottom-up saliency-driven auditory attention in [9], which could successfully detect pitch accents in English. The auditory attention model is biologically inspired and mimics the processing stages in the human auditory system. A set of multi-scale features is extracted from the sound spectrum based on the processing stages in the central auditory system and converted to low-level auditory gist features. Due to multi-scale structure of the model, auditory gist features capture rich information and can successfully recognize tones and pitch accents without requiring explicit prosodic feature, including pitch, computation.

The rest of the paper is organized as follows. The database used for experiments is described in Section 2. The auditory attention model together with gist extraction is explained in Section 3, which is followed by experimental results in Section 4. The concluding remarks are presented in Section 5.

2. DATABASE

The Boston University Radio News Corpus (BURNC) [10] was used for the experiments in English. The BURNC is a broadcast news-style read speech corpus that consists of speech from 3 female and 3 male speakers, totaling about

Table 1. Distribution of Pitch Accent and Tone Labels

Corpus	H*	!H*	L*	L+H*
BURNC	24.7%	54.4%	4.0%	16.8%

Corpus	L-H%	L-L%
BURNC	39.5%	60.5%

Corpus	Tone1	Tone2	Tone3	Tone4
MC-CC	20.2%	17.3%	22.9%	39.6%

3 hours of acoustic data, with ToBI-style [11] pitch accent and boundary tone annotations. Based on the distribution of pitch accent and boundary tone labels, here we used the most common four pitch accent categories, namely H* (high), !H* (down-stepped), L* (low), and L+H* (rising peak), and two boundary tone categories, namely L-L% and L-H% (low phrase accent followed by a low/or high boundary tone). The remaining categories formed an insignificant fraction of the corpus and were discarded. Approximately, 14.7K words carried one of these four pitch accent types and 5.6K of the words carried a boundary tone label. The distribution of pitch accent and boundary tone labels is presented in Table 1. The chance level, obtained by labeling all samples with the majority class, is 54.4% and 60.5% accuracy for pitch accent and boundary tone classification tasks, respectively.

Lexical tone recognition experiments are conducted using a continuous Mandarin Chinese speech database (referred to as MC-CC) that contains 7513 command-and-control utterances from 8 female and 8 male speakers. In Mandarin Chinese, each Chinese character represents a monosyllable and has one of the five tones; high-level (tone 1), high-rising (tone 2), low-dipping (tone 3), high-falling (tone 4), and neutral (tone 5). The neutral tone usually does not have a stable pitch contour and forms an insignificant fraction of the database; hence, it's ignored here. In the database, there are approximately 26K syllables which carry one of the four tone categories. The distribution of the tone labels in the MC-CC is listed in Table 1. The chance level is 39.6% accuracy.

3. AUDITORY ATTENTION MODEL

The block diagram of the auditory attention model is shown in Fig 1. As stated earlier, the model is biologically inspired and hence mimics the processing stages in the human auditory system. First, the auditory spectrum of the input sound is computed based on early stages of the human auditory system. The early auditory system model used here consists of cochlear filtering, inner hair cell, and lateral inhibitory stages mimicking the process from basilar membrane to the cochlear nucleus in the auditory system [9]. The cochlear filtering is implemented using a bank of 128 overlapping constant-Q asymmetric band-pass filters. For analysis, audio frames of 20 milliseconds (ms) with 10 ms shift are used, i.e. each 10 ms audio frame is represented by a 128 dimensional vector.

Next, multi-scale features, which consist of *intensity* (I), *frequency contrast* (F), *temporal contrast* (T), and *orientation* (O_θ) with $\theta = \{45^o, 135^o\}$, are extracted from the auditory spectrum based on the processing stages in the central

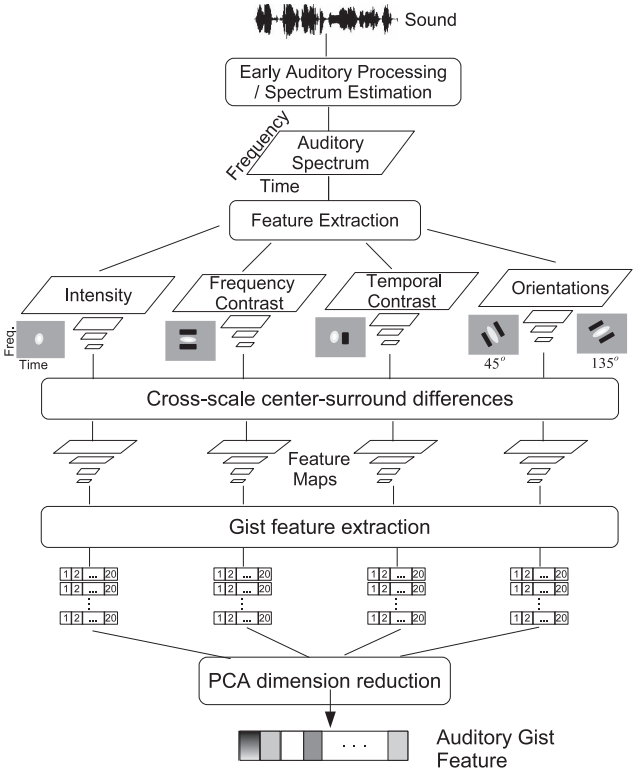

Fig. 1. Auditory Attention Model and Gist Extraction

auditory system [9, 12]. These features are extracted using 2D spectro-temporal receptive filters mimicking the analysis stages in the primary auditory cortex. Each of the receptive filters (RF) simulated for feature extraction is illustrated with gray scaled images in Fig 1 next to its corresponding feature. The excitation phase and inhibition phase are shown with white and black color, respectively. For example, the frequency contrast filter corresponds to receptive fields in the primary auditory cortex with an excitatory phase and simultaneous symmetric inhibitory side bands. Each of these filters is capable of detecting and capturing certain changes in signal characteristics. For example, the orientation features are capable of detecting and capturing when pitch is raising (orientation with 45^o) or falling (orientation with 135^o) [12].

The RF for generating frequency contrast, temporal contrast and orientation features are implemented using 2D Gabor filters with angles 0^o, 90^o, $\{45^o, 135^o\}$, respectively. The RF for intensity feature is implemented using a 2D Gaussian kernel. The multi-scale features are obtained using a dyadic pyramid: the input spectrum is filtered and decimated by a factor of two, and this is repeated. Finally, eight scales are created (if the scene duration is larger than 1.28 seconds (s); otherwise there are fewer scales), yielding size reduction factors ranging from 1:1 (scale 1) to 1:128 (scale 8). For details of the feature extraction and filters, one may refer to [9, 12].

After multi-scale features are obtained, the model computes "center-surround" differences by comparing "center"

fine scales with "surround" coarser scales yielding feature maps. The center-surround operation mimics the properties of local cortical inhibition and detects local temporal and spatial discontinuities. It is simulated by across scale subtraction ($\ominus$) between a center scale c and a surround scale s yielding a feature map $\mathcal{M}(c, s)$:

$$\mathcal{M}(c, s) = |\mathcal{M}(c) \ominus \mathcal{M}(s)|, \;\; \mathcal{M}\epsilon\{I, F, T, O_\theta\} \quad (1)$$

The across scale subtraction between two scales is computed by interpolation to the finer scale and point-wise subtraction. Here, $c = \{2, 3, 4\}$, $s = c + \delta$ with $\delta\epsilon\{3, 4\}$ are used, which results in 30 feature maps when there are eight scales.

Next, an "auditory gist" vector is extracted from the feature maps of I, F, T,O_θ such that it covers the whole scene at low resolution. To do that, each feature map is divided into m-by-n grid of sub-regions and mean of each sub-region is computed to capture the overall properties of the map. For a feature map $\mathcal{M}_i$ with height h and width w, the computation of feature can be written as:

$$G_i^{k,l} = \frac{mn}{wh} \sum_{u=\frac{kw}{n}}^{\frac{(k+1)w}{n}-1} \sum_{v=\frac{lh}{m}}^{\frac{(l+1)h}{m}-1} \mathcal{M}_i(u, v), \quad (2)$$

for $k = \{0, \cdots, n-1\}, l = \{0, \cdots, m-1\}$. An example of gist feature extraction with $m = 4, n = 5$ is shown in Fig 1, where a $4 \times 5 = 20$ dimensional vector is shown to represent a feature map. After extracting a gist vector from each feature map, we obtain the cumulative gist vector by augmenting them. Then, principal component analysis (PCA) is used to remove redundancy and to reduce the dimension to make machine learning more practical.

4. EXPERIMENTS AND RESULTS

In all the experiments, a 3-layer neural network is used for learning the mapping between the auditory gist features and tone and pitch accent classes, since it's biologically well motivated. The neural network has D inputs, $(D + N)/2$ hidden nodes and N output nodes, where D is the length of auditory gist vector after PCA dimension reduction when 95% of the variance is retained, and N is the number of classes. All of the results presented here are estimated using the average of 10-fold cross-validation by splitting data randomly into 80% training, 10% held-out, and 10% test sets.

First, pitch accent and boundary tone classification experiments are presented using the BURNC. A window that centers on a word is used in order to extract sound around the word. Then, auditory gist features are extracted from these sound segments for pitch accent and boundary tone classification. The mean duration of accented words in the BURNC is approximately 0.4 s. The role of window duration W is investigated in the experiments by varying duration from 0.2 s up to 1.4 s to analyze the effect of neighboring left and right context on the performance. Different grid sizes are tested for auditory gist extraction, and it is found that segmenting feature maps into 4-by-5 grids is sufficient and performs well. In Table 2, pitch accent and boundary tone classification results for varying window duration are presented together with the auditory gist dimension D. The pitch accent classification accuracy is lower for short scene durations, especially for the case when $W \leq 0.4$ s; i.e. when segment approximately includes only the word by itself. This essentially indicates that the co-articulation effect is significant and hence context information is important in pitch accent classification. As seen in Table 2, the best performance achieved is 64.6% pitch accent accuracy when $W = 1.2$ s and 89.7% boundary tone accuracy when $W = 0.6$ s.

The results compare well against the previously reported performance levels with the BURNC. We cannot compare our results directly to the ones in [1][2][3][5], due to different test conditions, including amount of data, number of speakers and pitch accent and boundary tone classes used, and hence the chance levels. However, we can compare our results directly with [4] since we used the complete BURNC data set as they did and have the same classes and chance levels in the experiments. [4] used RFC parameterization of pitch contour based on a neural network classifier together with a word language model and achieved 56.4% pitch accent accuracy and 67.7% boundary tone accuracy, which are presented as "Baseline-2" in Table 4. In summary, the proposed auditory attention cues provide 10.2% improvement over the chance level and 8.2% improvement over [4] on pitch accent classification. It also provides 29.2% improvement over chance level and 22% improvement over [4] on boundary tone classification.

Next, we tested proposed method on lexical tone recognition in Mandarin using the MC-CC dataset. In Mandarin, the final part of a syllable is regarded as voiced, whereas the initial part is regarded as consonant. In [8], it was found that using whole syllable features performs better than using features only from the final voiced part of a syllable. Thus, as in BURNC, a window that centers on a syllable is used in order to extract sound around the syllable for auditory gist feature computation. To obtain syllable time boundaries, the recorded speech was force aligned to the reference transcriptions. From the forced alignment output, it is found that the mean syllable duration is 0.28 s. The role of window duration W is investigated in the experiments by varying it from 0.2 s up to 1.2 s to analyze the co-articulation effect. The lexical tone recognition performance as a function of window duration is presented in Table 3. The tone recognition performance is poor for short scene durations; i.e. $W \leq 0.3$ s, when segment approximately includes only the syllable by itself. As before, this essentially indicates that the co-articulation of tones is significant and hence context information is important in lexical tone recognition. As seen in Table 3, the best performance achieved is 79.0% tone classification accuracy when $W = 0.8$ s.

We also compared our results with [8] which used prosodic features for tone recognition in Mandarin. As in [8], prosodic features consisting of pitch and duration are extracted from the initial (consonant) and final part (vowel) of a

Table 2. Pitch Accent and Boundary Tone Classification Results for BURNC for Varying Window Duration W

W (s)	Accent		Boundary	
	D	Accuracy	D	Accuracy
0.2	24	57.1%	19	87.5%
0.4	47	59.9%	39	89.1%
0.5	49	61.1%	42	89.4%
0.6	54	62.6%	**46**	**89.7%**
0.8	82	63.2%	70	88.4%
1.0	81	64.0%	71	87.8%
1.2	**87**	**64.6%**	78	87.1%
1.4	94	64.3%		
1.6	95	63.3%		

Table 3. Lexical Tone Classification Performance for MC-CC for Varying Window Duration W

W (s)	D	Accuracy
0.2	22	66.2%
0.3	30	72.8%
0.4	47	77.2%
0.5	49	78.3%
0.6	53	78.2%
0.8	**77**	**79.0**%
1.0	74	76.7%
1.2	76	75.8%

syllable. Pitch contour is smoothed, normalized and sampled to a fixed number of points. Also, the features that belong to the previous syllable are augmented in order to consider the left context. These context dependent features are normalized per speaker, and the dimension of the final feature vector was 14. Then, on MC-CC, 59.1% tone classification accuracy, which is presented as "Baseline-2" in Table 4, is obtained with these context dependent prosodic features. As seen in Table 4, the proposed method provides approximately 40% absolute improvement over the chance level and 20% absolute improvement over [8].

Finally, the contribution of each feature in the attention model is presented in Table 5 for the best performing W. All of the features individually perform well above the chance level for all the tasks. The most informative feature about pitch accent and tone classes was the orientation, O_θ with $\theta = \{45^o, 135^o\}$. However, the best performance is achieved with combined features $IFTO$.

5. CONCLUSION AND FUTURE WORK

In this paper, biologically inspired auditory attention cues are proposed for pitch accent and tone recognition. A set of multi-scale auditory features is extracted in parallel from the auditory spectrum of the sound and converted into low-level auditory gist features that capture the essence of sound. Using a neural network, the model learns the mapping between the gist features and pitch accent and tone classes. The model achieves 64.6% pitch accent and 89.7% boundary tone classification accuracy using the BURNC. These results compare very well with the reported human performance given that the agreement for manual annotators was around 60% for pitch accent classes and 90% for boundary tones for the BURNC [10]. In addition, the model was successful in tone recognition in Mandarin. The proposed method achieved 79% lexical tone accuracy and provided 20% absolute improvement over a state-of-the art method with traditional prosodic features.

As part of our future work, we plan to conduct experiments in other languages. Also, we plan to include tone information obtained from the model into an automatic speech recognition system to improve speech recognition performance in tone languages.

Table 4. Summary and Comparison of Results

Method	BURNC		MC-CC
	Accent	Boundary	Tone
Baseline-1 (chance)	54.4%	60.5%	39.6%
Baseline-2 (prosodic)	56.4%	67.7%	59.1%
Attention Cues	**64.6%**	**89.7%**	**79.0%**

Table 5. Accuracy(%) Achieved with Individual Feature

Corpus	Task	I	F	T	O	$IFTO$
BURNC	Accent	57.1	61.5	58.5	62.0	64.6
	Boundary	81.2	80.7	74.4	87.7	89.7
MC-CC	Tone	52.2	57.8	55.0	74.1	79.0

6. ACKNOWLEDGEMENT

The author would like to thank Dr. Ruxin Chen and Dr. Xavier Menendez-Pidal at SCEA for helping with Mandarin force alignment experiments.

7. REFERENCES

[1] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, 1996.
[2] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proc. of ICSLP*, 2002.
[3] G.A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *Proc. of HLT*, 2006.
[4] S. Ananthakrishnan and S. Narayanan, "Fine-grained pitch accent and boundary tone labeling with parametric f0 features," in *Proc. of ICASSP*, 2008.
[5] A. Rosenberg, "Classification of prosodic events using Quantized Contour Modeling," in *Proc. of HLT*, 2010.
[6] C.J. Chen, R.A. Gopinath, M.D. Monkowski, M.A. Picheny, and K. Shen, "New methods in continuous Mandarin speech recognition," in *Eur. Conf. on Speech Comm. Tech.*, 1997.
[7] C. Wang, *Prosodic Modeling for Improved Speech Recognition and UnderStanding*, Ph.D. thesis, MIT, 2001.
[8] X. Lei, M. Siu, M. Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for mandarin broadcast news speech recognition," in *Proc. of Interspeech*, 2006.
[9] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. of Interspeech*, 2007.
[10] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, *The Boston University Radio Corpus*, 1995.
[11] K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg, "ToBI: A standard scheme for labeling prosody," in *Proc. of ICSLP*, 1992.
[12] O. Kalinli and S. Narayanan, "Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information," *IEEE Trans. Audio, Speech, Lang. Process.*, 2009.