Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Exploiting articulatory features for pitch accent detection*

Junhong ZHAO[†1,2], Ji XU[3], Wei-qiang ZHANG[3], Hua YUAN[3], Jia LIU[3], Shanhong XIA[1]

([1]*State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China*)

([2]*University of Chinese Academy of Sciences, Beijing 100190, China*)

([3]*National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*)

[†]E-mail: junhong.iecas@gmail.com

Received Apr. 22, 2013; Revision accepted Sept. 29, 2013; Crosschecked Oct. 15, 2013

Abstract: Articulatory features describe how articulators are involved in making sounds. Speakers often use a more exaggerated way to pronounce accented phonemes, so articulatory features can be helpful in pitch accent detection. Instead of using the actual articulatory features obtained by direct measurement of articulators, we use the posterior probabilities produced by multi-layer perceptrons (MLPs) as articulatory features. The inputs of MLPs are frame-level acoustic features pre-processed using the split temporal context-2 (STC-2) approach. The outputs are the posterior probabilities of a set of articulatory attributes. These posterior probabilities are averaged piecewise within the range of syllables and eventually act as syllable-level articulatory features. This work is the first to introduce articulatory features into pitch accent detection. Using the articulatory features extracted in this way, together with other traditional acoustic features, can improve the accuracy of pitch accent detection by about 2%.

Key words: Articulatory features, Pitch accent detection, Prosody, Computer-aided language learning (CALL), Multi-layer perceptron (MLP)

doi:10.1631/jzus.C1300104 **Document code:** A **CLC number:** TP391; TN912.34

1 Introduction

Prosody, as the supra-segmental information of speech, plays a critical role in stress-timed languages such as English. It manifests itself in many aspects such as accent, pause, and intonation. Among them, the most essential aspect is accent. It always goes with amplified energy, long duration, extreme variation of fundamental frequency, and good pronunciation quality of speech.

The mastery of pronunciation skills using accent can help speakers to express their key point, purpose, attitude, and so on. However, in secondary language acquisition, it seems more difficult for learners to master accentual than phonetic pronunciation for several reasons (Meng *et al*., 2009). Sometimes, even if every phoneme is pronounced correctly, the sentence still sounds non-native or unnatural if the expressions of accent are not good enough. Thus, for the computer-aided language learning (CALL) system, module incorporation of pitch accent detection will be beneficial since it can feed back speakers' mispronunciations at a prosodic level to help speakers master languages better.

Up until now, many approaches have been investigated for pitch accent detection. These approaches use various prosodic features extracted within the range of syllables. Some of these features are extracted in a statistical way, for example, the maximum, minimum, or mean value and the variation range of the pitch or energy. The others are extracted based on the contour of the pitch or energy, such as the coefficients of polynomial expansion and the upsample values. To model these features, various classifiers have been used, like support vector machines (SVMs) (Jeon and Liu, 2009a; 2009b; 2010; 2012),

* Project (Nos. 61370034, 61273268, and 61005019) supported by the National Natural Science Foundation of China

artificial neural networks (Ananthakrishnan and Narayanan, 2008), and decision trees (Sun, 2002).

Pitch accent not only gives prominence to suprasegmental features such as duration, pitch, and energy, but also brings differences to articulatory variation from the perspective of speech production. When vocalizing, a speaker will adjust the shape of vocal tract through moving articulators such as the lips, tongue, palate, and teeth. The information on the movement of these articulators, including the manner of articulation (such as nasal and fricative) and the place of articulation (such as back, middle, or front position of the tongue), is known as articulatory features (AFs) (Iribe *et al*., 2010). In general, one phoneme often has multiple AFs. For example, the phoneme /æ/ has such AFs as voiced, vowel, continuant, low position of tongue, and tense vocal cord. In this work, AFs will be represented by the posterior probabilities of various kinds of articulatory attributes.

When pronouncing accented syllables, speakers often make their vocal cords vibrate more tensely, the lips rounder, and the mouth opened more widely (Fougeron, 1999; Erickson, 2002; Cho, 2006). In the literature the relationship between pitch accent and the articulatory movement was studied either from the image of the vocal tract (Erickson, 2002) or articulatory kinematics data collected using electromagnetic midsagittal articulography (EMA) (Cho, 2006). It was shown consistently that pitch accent will introduce prominence of articulatory variation, especially for mouth shape. These distinctions in articulation make AFs useful in pitch accent detection. However, this has not yet been much explored in this field.

In summary, the following advantages have made AFs widely used in speech recognition: (1) AFs are not sensitive to environment changes; (2) AFs are robust for different speakers; (3) In a sense AFs are universal for different languages. These outstanding advantages have been well manifested by the success of many related studies devoted to speech recognition under different scenarios. Kirchhoff *et al*. (2002) exploited AFs for robust speech recognition and proved that AFs can be used to significantly improve the performance of speech recognition in low signal-to-noise ratio (SNR) environments. Siniscalchi *et al*. (2008) built a language-independent phone recognizer based on a bank of articulatory detectors. The recognizer can achieve good recognition performance across multiple languages. Because there is a universal set of articulatory attributes for several diverse languages and the corresponding training material can be shared, AFs are also used for low-resource speech recognition (Qian *et al*., 2011; 2013; Qian and Liu, 2012a; 2012b), where the training data used by the target language can be borrowed from close non-target language data.

For some other applications, AFs also play an important role and have been successfully used. In the field of CALL, Iribe *et al*. (2010; 2012) used AFs to evaluate the learner's pronunciation, and visualized them in real time using animation synthesis technology for mispronunciation correction. Sangwan *et al*. (2010) and Sangwan and Hansen (2012) used AFs to conduct language and accent identification, respectively, and achieved good performance for both tasks. Based on the impact of articulatory characteristics on fundamental frequency, Chao *et al*. (2012) effectively used AFs in tone recognition of Mandarin. By combining AFs with other prosodic features, the tone recognition accuracy was significantly improved. Recently, more and more attention has been paid to AFs in expressive speech synthesis, as a competitive parameterization of the speech signal (Black *et al*., 2012).

In this paper, we investigate the benefits of employing AFs to detect pitch accent in English. Several approaches have been proposed on the extraction of AFs, such as direct measurement by cine-radiography (Papcun *et al*., 1992) and transformation from the acoustic signal by filtering (Schroeter and Sondhi, 1994; Richards *et al*., 1996; 1997; Krstulovic, 1999). The most commonly used approach is the data-driven method, in which AFs are the posterior probabilities of articulatory attributes produced by statistical classification. Also, it is the approach that we use in this study. Specifically, we refer to the extraction method used by Siniscalchi *et al*. (2008) and Qian and Liu (2012a) to extract the AFs. The acoustic features are first extracted and pre-processed. Then multi-layer perceptron (MLP) models of articulatory attributes are trained. Through these models, the posterior probabilities of articulatory attributes can be obtained. We use the same framework and structure of two-layer MLPs as used in Siniscalchi *et al*. (2008) and Qian and Liu (2012a) to model articulatory attributes, but we have adjusted the choices of articulatory attributes for each layer of MLP according to our special task of pitch accent. To obtain the syllable-

level AFs that can be used for detecting pitch accent, the frame-level probabilities are averaged piecewise within the syllable range. Our experiments prove that AFs extracted in this way can be used to effectively improve the accuracy of pitch accent detection, in both binary and four-way detection tasks.

2 Feature extraction

2.1 Articulatory feature extraction

The approach we adopt here to extract AFs consists of three steps: (1) acoustic feature pre-processing, (2) articulatory attribute modeling and decoding, and (3) post-processing.

2.1.1 Acoustic feature pre-processing

The split temporal context-2 (STC-2) approach (Schwarz *et al*., 2006) is used as the front end of our system to pre-process acoustic features. In the STC-2 approach, the frame-based acoustic features (perceptual linear predictive (PLP) features here) are first extended within a block to take advantage of contextual dependencies, by concatenating the acoustic features of preceding and following frames with the current frame itself. The block spans 31 frames in this work, and is split into two groups, frames 1–16 and 16–31, forming two data streams which will be processed separately afterwards. The PLP features from these two groups are both adjusted using hamming window and transformed to the discrete cosine transform (DCT) domain, to de-correlate and reduce dimensionality.

2.1.2 Articulatory attribute modeling and decoding

After pre-processing, the two data streams of acoustic features will be input into hierarchical MLP networks with the divide-and-merge architecture (Fig. 1). The networks in different layers are organized in different ways. In the first layer, there are many small scale networks, called AF-detector MLPs here. Each AF-detector MLP corresponds to one articulatory attribute. These MLPs are independent of each other. Here we use the set of articulatory attributes proposed by Siniscalchi *et al*. (2008) and add two extra categories, unaccented vowels and accented vowels, since we treat accented and unaccented vowels separately in our phoneme set. All the mapping relationships between attributes and phonemes are listed in Table A1 in the Appendix.

For preparation of the transcription needed for training, the phonetic transcription (43 phonemes in all) is first generated by forced alignment, using the speaker-independent speech recognizer pre-trained with all the training data. Then the accented labels are integrated according to prosodic manual annotation. Based on the transcription, we label the training data into attribute present or attribute absent region for every articulatory event. This process is completed automatically by transferring available phoneme-level transcription to the attribute level according to the mapping relationships between attributes and phonemes listed in Table A1 in the Appendix. For each AF-detector MLP network, two output nodes are configured to model each of these two kinds of attribute data. For modeling attributes more elaborately, we let each output node have three states, which are used to model the initial, middle, and ending parts of the attribute, respectively. The timing information used for training is obtained from the state-level forced alignment of the corresponding phonemes.

In the second layer, only one large-scale network is used to merge all probabilities output by separated AF-detector MLPs. Here we call this network 'merger MLP'. For the choice of merger events used in this layer, as pitch accent acts on the vowel part of the

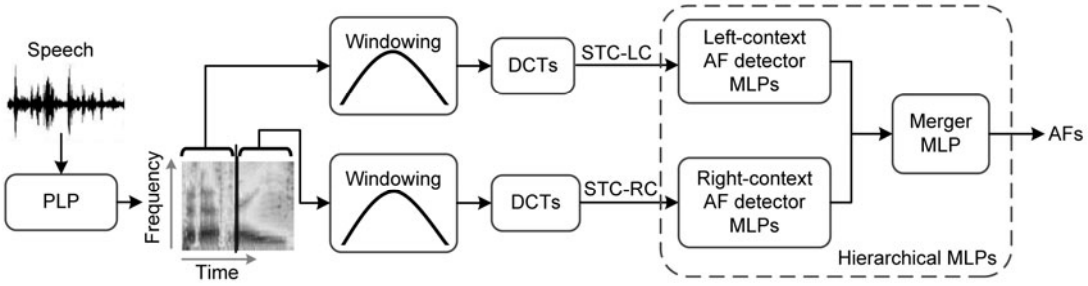

Fig. 1 The pipeline of articulatory feature (AF) extraction

word, we focus on the description of vowels. According to the five vowel letters and their possible pronunciations, we select five merger events for each of the accented and unaccented categories. For consonant phonemes, four typical categories are chosen, including fricative, nasal, stop, and approximant. Thus, there are 15 merger events (Table 1). This is different from Siniscalchi *et al*. (2008), who used phonemes as merger events for phone classification and recognition.

Table 1 The look-up table of the categories used in the merger layer and their corresponding phonemes

Category for merger MLP	International phonetic alphabet
Fricative	/dʒ/, /tʃ/, /s/, /ʃ/, /z/, /f/, /θ/, /v/, /ð/, /h/, /ʒ/
Nasal	/m/, /n/, /ŋ/
Stop	/b/, /d/, /g/, /p/, /t/, /k/
Approximant	/w/, /j/, /l/, /r/
U-IVowel	/i/, /i:/
U-EVowel	/e/, /ei/
U-AVowel	/ɔ/, /æ/, /au/, /ai/, /ə:/
U-OVowel	/ɔ:/, /əu/, /ɔi/
U-UVowel	/ʌ/, /ə/, /əl/, /əm/, /ən/, /u/, /u:/
A-IVowel	/i'/, /i:'/
A-EVowel	/e'/, /ei'/
A-AVowel	/ɔ'/, /æ'/, /au'/, /ai'/, /ə:'/
A-OVowel	/ɔ:'/, /əu'/, /ɔi'/
A-UVowel	/ʌ'/, /ə'/, /u'/, /u:'/
Silence	Pauses

The detailed organization of the hierarchical MLPs is depicted in Fig. 2. There are 23 parallel MLPs in the first layer. The data stream from the left or right context is modeled using the respective group of MLPs. Every AF-detector MLP in the same group receives the same input of acoustic features but models its own articulatory attribute. There are two output nodes, each with three states, for each AF detector to represent how likely the input frame possesses the articulatory attribute. Combining all of these probabilities we obtain the LC-AF vector and RC-AF vector. These two AF vectors are concatenated together as the concatenated-AF vector and input into the merger layer. From the output of the merger layer, we obtain the merged-AF vectors.

In this study, all MLPs are three-layer networks (one input layer, one hidden layer, and one output layer) with fully connected structure. There are 500 hidden units for the AF detector MLP while 1500 for the merger MLP. The ICSI QuickNet software package for neural network ICSI (http://www.icsi.berkeley.edu/Speech/qn.html) is employed to build these neural networks, using classical back propagation to minimize the cross entropy error between outputs and targets. The learning rate and stopping criterion are controlled by the frame-based classification error rate on the cross validation data set.

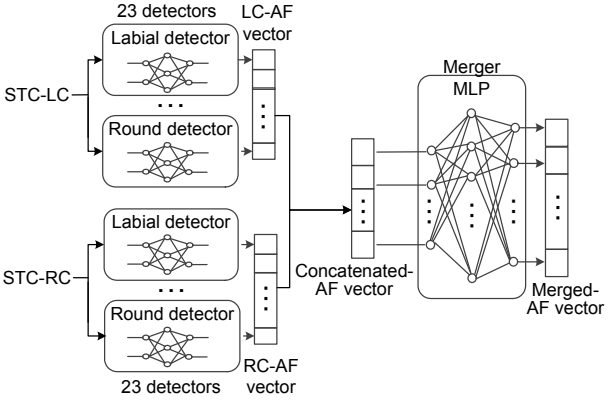

Fig. 2 The structure of hierarchical multi-layer perceptrons (MLPs)

For the hierarchical MLPs used to extract AFs, the computational complexities of the two layers are analyzed here. In the first layer, there are 23 AF-detector MLPs ($d^1_{\text{detector}}=23$). For each detector, the dimensionality of input features is $d^1_{\text{fea}}=16\times 39$ (16 frames, 39 dimensions of PLP features for each frame) and there are two input streams (LC and RC, $d^1_{\text{stream}}=2$), so the dimensionality of inputs for each detector is $d^1_{\text{input}}=d^1_{\text{fea}}d^1_{\text{stream}}$. Each AF detector has two output nodes with three states, so the dimensionality of outputs for each detector is $d^1_{\text{output}}=d^1_{\text{stream}}\times 2\times 3$. There are 500 hidden units ($d^1_{\text{hidden}}=500$) in the hidden layer for each detector. Thus, the overall complexity of the first-layer networks is $O((d^1_{\text{input}}d^1_{\text{hidden}}+d^1_{\text{hidden}}d^1_{\text{output}})d^1_{\text{detector}})$. In the second layer, there is only one MLP. Its inputs are the concatenated-AF vectors with $d^1_{\text{detector}}d^1_{\text{output}}$ dimensions. The outputs are 15 nodes corresponding to 15 categories of merger events; also, there are three states for each node. Thus, the dimensionality of outputs is $d^2_{\text{detector}}=15\times 3$. There are 1500 hidden units in the

hidden layer ($d^2_{\text{hidden}}=1500$). Thus, the complexity of the second-layer network is $O(d^1_{\text{detector}}d^1_{\text{output}}d^2_{\text{hidden}}+d^2_{\text{hidden}}d^2_{\text{output}})$.

2.1.3 Feature mapping from the frame level to syllable level

The AFs output from MLPs are all at the frame level. As Fig. 3 illustrates, the AF vector of the *i*th frame is denoted by $\boldsymbol{X}_i$. To obtain syllable-level AFs for use in pitch accent detection, we average these frame-level features within the syllable scope.

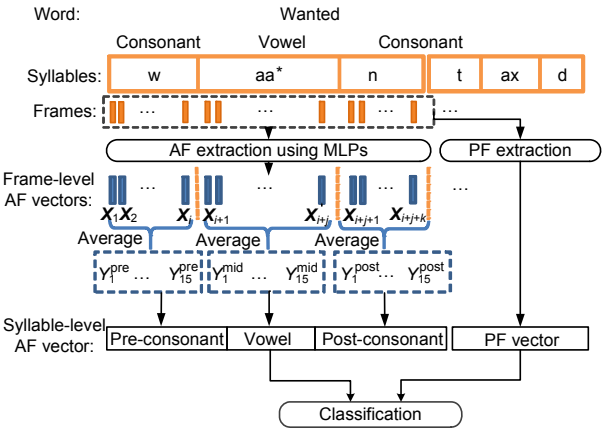

Fig. 3 Illustration of feature mapping from the frame level to the syllable level

Based on the recognized phoneme sequence and their boundary information, syllable parse will be conducted in advance to obtain syllable divisions for both training and test data. The boundary information of syllables is decided by the starting time of its first phoneme and ending time of its last phoneme.

There are several consonants and one vowel in one syllable. The useful information lies in not only the vowel but also the consonants. We divide the syllable into three parts, pre-consonant, vowel, and post-consonant, according to general syllable structure, and perform the averaging in these three regions respectively. Multiple preceding (following) consonants will be put into the same part. The averaged syllable-level AFs will then be used by the classifier to improve the performance of pitch accent detection. Details have been illustrated in Fig. 3.

2.2 Extracting traditional prosodic features

Besides the averaged syllable-level AFs extracted using the method mentioned above, many other traditional prosodic features are extracted in our system. These prosodic features are related to energy, pitch, and duration. There are 24 features in all, which can be summarized as the following three categories:

Frame-averaged features (4): loudness, semitone, spectral emphasis, and duration. For perceptual accuracy, here we use loudness to replace energy and semitone to replace pitch values in Hz to detect pitch accent. Both of loudness and semitone are extracted following our method proposed in Zhao *et al*. (2011). The spectral emphasis, which is the mid-frequency energy in the range of 500–2000 Hz, has been reported to be more powerful in pitch accent classification than full-range energy (Sluijter and van Heuven, 1996) and thus is used in this study. To reduce the negative impact brought by inter-speaker variation and speech rate differences, all the four features are normalized by the mean value of the whole sentence.

TILT features (4): Our work follows the rise/fall/connection (RFC) intonational model proposed by Taylor (1994), and the TILT parameter (Taylor, 1998) set is extracted to describe the variation of the pitch contour.

Differential features (16): Considering the impact of co-articulation and prosodic context, we also extract the forward and backward differences for both the frame-averaged basic features and TILT features.

3 Experiments

3.1 Data corpus and experimental setup

The experiments are performed on the Boston University Radio Speech Corpus (BURSC) (Ostendorf *et al*., 1995). This corpus consists of broadcast news style speech read by seven radio news announcers. Most of the data is hand-annotated with prosodic information based on the ToBI labeling architecture. In the experiments, we use 420 paragraphed-size utterances from six speakers (F1A, F2B, F3A, M1B, M2B, M3B). The distributions of data are listed in Table 2. We randomly arrange these utterances and perform five-fold cross-validations for all the experiments.

We devise two tasks, two-way and four-way, to detect pitch accent. In the two-way detection task, we just investigate the binary distribution of accented and

unaccented syllables. In the transcription preparation for training, once there is a suffix mark '*' in the primitive prosodic label, the syllable is treated as accented and the rest is treated as unaccented. In the four-way detection task, we finely partition accented types into three sub-classes: high, low, and down-stepped. In transcription preparation, the syllables are allocated as these three classes when 'H*', 'L*', and '!' markers exist in the labels. The three uncertainty labels '*', '*?', and 'X*?' are taken as high pitch accent markers here. Through these two tasks, we can investigate the effectiveness of using AFs under different conditions.

Table 2 The distributions of data used in our experiment

Speaker	Number of utterances	Number of sentences	Number of words	Number of syllables	Number of accents
F1	74	271	3993	7234	2253
F2	166	1176	12691	23199	7063
F3	33	160	2733	4680	1530
M1	71	391	5059	8723	2564
M2	50	209	3608	6288	1933
M3	24	122	2094	3527	1013

3.2 Recognition accuracy

The performance of each articulatory attribute model in the AF detector layer is evaluated by frame-level recognition accuracy with respect to the left and right context of the testing data (Fig. 4). For each articulatory attribute model, the recognition performances for the data from the left context and right context are nearly the same. The frame-level accuracy of most AFs is acceptable; only the recognition accuracy of Dental and Glottal is relatively low, perhaps owing to the sparseness of the training data. As noted in the mapping relationship between attributes and phonemes in Table A1, only the phoneme /h/ has the Dental attributes. Similarly, only two phonemes /ð/ and /θ/ have the Glottal attributes. The fewer the phonemes one articulatory attribute corresponds to, the less the training data it can use. This will make the attribute model under-trained.

3.3 Comparison of different classification methods

Using the classification information provided by merger MLP, various criteria can be used to decide the final classification result. In this study, we use three methods, maximum and summation-greater classification principles and SVM, to determine the final classification results. Both the maximum and summation-greater principles are applied to the vowel part of the syllable-level AF vectors, while SVM is applied to the entire vectors, including pre- and post-consonant.

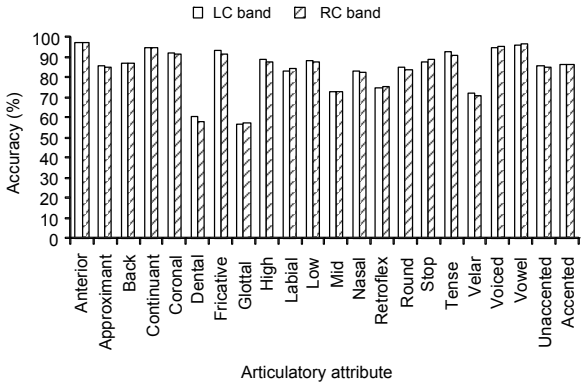

Fig. 4 Frame-level recognition accuracy of articulatory feature (AF) detectors

For the maximum criterion, when the articulatory attribute with the maximum probability belongs to the accented category, the syllable is detected as pitch accented. For the summation-greater principle, if the sum of the probabilities of all accented articulatory attributes is greater than that of unaccented ones, the syllable will be detected as accented. This can be denoted as Eqs. (1) and (2):

Maximum criterion:

$$\begin{aligned}&\text{If } \operatorname{argmax}_S P(S|O) \in \Phi \rightarrow \text{accented};\\ &\text{If } \operatorname{argmax}_S P(S|O) \notin \Phi \rightarrow \text{unaccented}.\end{aligned} \tag{1}$$

Summation-greater criterion:

$$\begin{aligned}&\text{If } \sum_{S_i \in \Phi} P(S_i|O) > \sum_{S_j \in \Psi} P(S_j|O) \rightarrow \text{accented};\\ &\text{If } \sum_{S_i \in \Phi} P(S_i|O) \le \sum_{S_j \in \Psi} P(S_j|O) \rightarrow \text{unaccented}.\end{aligned} \tag{2}$$

Φ and Ψ refer to the collections of accented and unaccented articulatory attributes respectively, and $P(S|O)$ indicates the probabilities vector. Beside these two criteria, the SVM method is used to find the classification hyperplane of the whole data set. Here, the SVM model is trained using LIBSVM (Fan *et al*.,

2005).

The performances of these three classification methods on two-way pitch accent detection are shown in Table 3. The accuracy achieved using the SVM approach is the highest. Although the complexity of SVM is relatively large ($O(n^2)$ vs. $O(n)$), it can handle AFs and other prosodic features in a consistent way, and thus is beneficial to feature combination. Therefore, we adopt the SVM method in the following experiments.

Table 3 Performance of articulatory features with different classification methods on two-way pitch accent detection

Classification method	Recognition accuracy (%)
Maximum	78.45
Summation-greater	78.75
SVM	79.98

3.4 Performance of articulatory features in pitch accent detection

We conduct two experiments to investigate the performance of AFs in pitch accent detection. In the first experiment, we adopt the MLP structure with only the merger layer to extract probability features. In the second experiment, we adopt the hierarchical MLP structure with both the detector and merger layers to extract probability features. Our objective is to verify the effect of AFs by comparing the pitch accent detection performance of probability features extracted with or without using AF information. To maintain consistency between the two experiments, the single-layer MLP is also trained using STC-2 features. Fig. 5 shows the frame-level recognition accuracy of merger events modeled by these two MLP structures.

The probability features extracted are used in two-way pitch accent detection with SVM as the classifier. As shown in Table 4, the system performance with hierarchical architecture is superior to that with a single layer. The probability features extracted by the hierarchical architecture are more powerful, not only when used alone but also when combined with prosodic features. This indicates that AFs provide useful distinctive information for pitch accent and can effectively improve the accuracy of pitch accent detection.

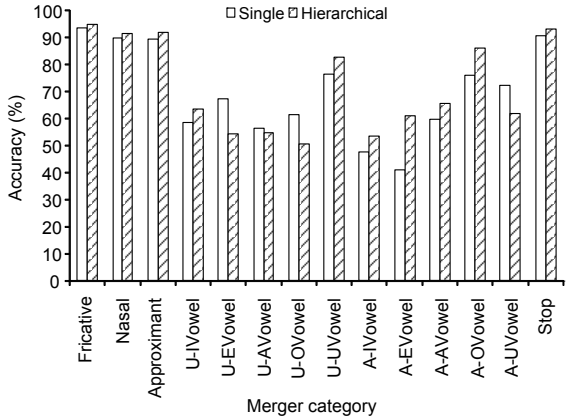

Fig. 5 Frame-level recognition accuracy of the merger event obtained using different MLP structures

Table 4 Performance of articulatory features extracted using single or hierarchical MLP structure on two-way pitch accent detection

MLP structure	Recognition accuracy (%)	
	Merged-AF	Merged-AF & PF
Single	78.79	79.95
Hierarchical	79.98	81.20

AF: articulatory feature; PF: probability feature

3.5 Combining articulatory features and traditional prosodic features

To investigate the effectiveness of AFs in pitch accent detection under different conditions, we conduct experiments of two- and four-way pitch accent detection, respectively. We use the same AFs in two- and four-way tasks; by the four-way task, we aim to complete a more specific pitch accent classification. Table 5 shows that merged-AFs are effective in pitch accent detection in both two- and four-way tasks. They can be an alternative to traditional prosodic features. Combining the merged-AFs with traditional prosodic features, the detection accuracy can be improved by about 2% for the two-way task and by about 1.8% for the four-way task.

Table 5 Performance of different features for two- and four-way pitch accent detection

Task	Recognition accuracy (%)		
	PF	Merged-AF	PF & Merged-AF
Two-way	79.16	79.98	81.20
Four-way	75.00	74.98	76.84

AF: articulatory feature; PF: probability feature

3.6 Importance analysis

In this experiment, we make a detailed analysis of the AFs to determine which categories have a closer relationship with pitch accent classification. To investigate the importance of different AFs on two-way pitch accent classification, we first map the concatenated-AFs from the frame level to the syllable level. After mapping, there will be 36 feature dimensions (three sections for each syllable, two data streams of left and right, three states for each MLP output node, and two decision values for each state) belonging to each AF for each syllable. A correlation-based feature selection (Hall, 1999; Hall and Smith, 1999) is then implemented on the entire data set to rank all these features by their importance. The ranking is conducted using the WEKA machine learning toolkit (Witten and Frank, 2005).

Among the ranked results, the top 100 dimensions are analyzed. These dimensions mostly come from the vowel part of the syllable and across almost every state. We assume that the AFs that have more than three dimensions in this list are more distinctive in pitch accent detection. Table 6 lists these articulatory attributes and the corresponding number of total feature dimensions existing in the top list.

As expected, Accented-vowel and Unaccented-vowel are the most important attributes and have the strongest correlation with pitch accent detection. Several attributes that relate to the place of vowel articulation are closely correlated to pitch accent, including Tense which defines the place of articulation as "sounds requiring deliberate, accurate, maximally distinct gestures that involve considerable muscular effort", Round which defines the place of articulation as "the lips are protruded", and Mid and Low which define the place of articulation as "the body of the tongue is lowered from the neutral position". Among them, Mid and Low are the main attributes for the jaw. When a human is speaking, the palate almost remains still. It is the movement of the jaw that is relied on in adjusting the mouth opening. With the movements of lips protruding to adjust mouth rounding, the articulatory attributes of Mid, Low, and Round are three main features for describing the shape of the mouth. It can be concluded that the information from the shape of the mouth and the muscular effort of the vocal cord plays a more important role in discriminating pitch accent, which is consistent with what was declared in Fougeron (1999) and Cho (2006). Also, some attributes that typically characterize consonant articulation like Glottal, Dental, and Velar are very determinative for pitch accent detection, because the phonemes that have such attributes can obviously be classified in an unaccented category.

Table 6 The articulatory features ranked by importance

Articulatory attribute	Feature dimensionality
Accented-vowel	21
Unaccented-vowel	17
Tense	10
Round	4
Mid	9
Low	4
Glottal	10
Dental	6
Velar	4

4 Conclusions and future work

In this paper, articulatory features (AFs) are first explored to improve the accuracy of pitch accent detection. The articulatory attributes are modeled by MLPs. The posterior probabilities output from these models are considered as the AFs. Experiments showed that the AFs extracted in this way are effective for pitch accent detection, not only when used alone but also when combined with traditional prosodic features. They can achieve about 2% and 1.8% detection accuracy improvements in two- and four-way tasks, respectively. Among these AFs, the ones that describe the shape of the mouth (Round, Mid, Low) and the muscular effort of the vocal cord (Tense) have a closer relationship with the pitch accent phenomenon.

Obviously, the performance of using AFs can be further improved. As the movement of human organs is continuous in space, much information is lost in current binary format of AFs. Thus, it is necessary to develop more sophisticated methods to model the articulatory attributes, in order to extract more discriminative information for pitch accent detection. In further work, we will address these problems and find more distinctive AFs for pitch accent detection.

References

Ananthakrishnan, S., Narayanan, S., 2008. Automatic prosodic event detection using acoustic, lexical and syntactic evidence. *IEEE Trans. Audio Speech Lang. Process.*, **16**(1): 216-228. [doi:10.1109/TASL.2007.907570]

Black, A.W., Bunnell, H.T., Dou, Y., Muthukumar, P.K., Metze, F., Perry, D., Polzehl, T., Prahallad, K., Steidl, S., Vaughn, C., 2012. Articulatory Features for Expressive Speech Synthesis. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4005-4008. [doi:10.1109/ICASSP.2012.6288796]

Chao, H., Yang, Z.L., Liu, W.J., 2012. Improved Tone Modeling by Exploiting Articulatory Features for Mandarin Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4741-4744. [doi:10.1109/ICASSP.2012.6288978]

Cho, T., 2006. Manifestation of prosodic structure in articulatory variation: evidence from lip kinematics in English. *Lab. Phonol.*, **8**:519-548.

Erickson, D., 2002. Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, **59**(2-3):134-149. [doi:10.1159/000066067]

Fan, R.E., Chen, P.H., Lin, C.J., 2005. Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.*, **6**:1889-1918.

Fougeron, C., 1999. Prosodically Conditioned Articulatory Variations: a Review. UCLA Working Papers in Phonetics, p.1-74.

Hall, M.A., 1999. Correlation-Based Feature Selection for Machine Learning. PhD Thesis, The University of Waikato, New Zealand.

Hall, M.A., Smith, L.A., 1999. Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. Proc. 12th Int. Florida Artificial Intelligence Research Society Conf., p.235-239.

Iribe, Y., Mori, T., Katsurada, K., Nitta, T., 2010. Pronunciation Instruction Using CG Animation Based on Articulatory Features. Proc. Int. Conf. on Computers in Education, p.501-508.

Iribe, Y., Mori, T., Katsurada, K., Kawai, G., Nitta, T., 2012. Real-Time Visualization of English Pronunciation on an IPA Chart Based on Articulatory Feature Extraction. Proc. Interspeech, p.1271-1274.

Jeon, J.H., Liu, Y., 2009a. Automatic Prosodic Events Detection Using Syllable-Based Acoustic and Syntactic Features. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4565-4568. [doi:10.1109/ICASSP.2009.4960646]

Jeon, J.H., Liu, Y., 2009b. Semi-supervised Learning for Automatic Prosodic Event Detection Using Co-training Algorithm. Proc. ACL-IJCNLP, p.540-548. [doi:10.3115/1690219.1690222]

Jeon, J.H., Liu, Y., 2010. Syllable-Level Prominence Detection with Acoustic Evidence. Proc. Interspeech, p.1772-1775.

Jeon, J.H., Liu, Y., 2012. Automatic prosodic event detection using a novel labeling and selection method in co-training. *Speech Commun.*, **54**(3):445-458. [doi:10.1016/j.specom.2011.10.008]

Kirchhoff, K., Fink, G.A., Sagerer, G., 2002. Combining acoustic and articulatory feature information for robust speech recognition. *Speech Commun.*, **37**(3-4):303-319. [doi:10.1016/S0167-6393(01)00020-6]

Krstulovic, S., 1999. LPC-Based Inversion of the DRM Articulatory Model. Proc. European Conf. on Speech Communication and Technology, p.125-128.

Meng, H., Tseng, C.Y., Kondo, M., Harrison, A., Viscelgia, T., 2009. Studying L2 Suprasegmental Features in Asian Enlishes: a Position Paper. Proc. Interspeech, p.1715-1718.

Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S., 1995. The Boston University Radio News Corpus. Linguistic Data Consortium.

Papcun, J., Hochberg, T.R., Thomas, F., Larouche, J., Zacks, J., Levy, S., 1992. Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data. *J. Acoust. Soc. Am.*, **92**(2):688-700. [doi:10.1121/1.403994]

Qian, Y.M., Liu, J., 2012a. Articulatory Feature Based Multilingual MLPs for Low-Resource Speech Recognition. Proc. Interspeech, p.2602-2605.

Qian, Y.M., Liu, J., 2012b. Cross-Lingualand Ensemble MLPs Strategies for Low-Resource Speech Recognition. Proc. Interspeech, p.2582-2585.

Qian, Y.M., Povey, D., Liu, J., 2011. State-Level Data Borrowing for Low-Resource Speech Recognition Based on Subspace GMMs. Proc. Interspeech, p.553-560.

Qian, Y.M., Xu, J., Liu, J., 2013. Multi-stream posterior features and combining subspace GMMs for low resource LVCSR. *Chin. J. Electron.*, **22**(2):291-295.

Richards, H.B., Mason, J.S., Hunt, M., Bridle, J., 1996. Deriving Articulatory Representations of Speech with Various Excitation Modes. Proc. 4th Int. Conf. on Spoken Language, p.1233-1236. [doi:10.1109/ICSLP.1996.607831]

Richards, H.B., Bridle, J., Hunt, M., Mason, J.S., 1997. Vocal Tract Shape Trajectory Estimation Using MLP Analysis-by-Synthesis. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.1287-1290. [doi:10.1109/ICASSP.1997.596181]

Sangwan, A., Hansen, J.H.L., 2012. Automatic analysis of Mandarin accented English using phonological features. *Speech Commun.*, **54**(1):40-54. [doi:10.1016/j.specom.2011.06.003]

Sangwan, A., Mehrabani, M., Hansen, J.H.L., 2010. Automatic Language Analysis and Identification Based on Speech Production Knowledge. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.5006-5010. [doi:10.1109/ICASSP.2010.5495066]

Schroeter, J., Sondhi, M.M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech Audio Process.*, **2**(1):133-150. [doi:10.1109/89.260356]

Schwarz, P., Matejka, P., Cernocky, J., 2006. Hierarchical Structure of Neural Networks for Phoneme Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal

Processing, p.325-328. [doi:10.1109/ICASSP.2006.1660023]

Siniscalchi, S.M., Svendsen, T., Lee, C.H., 2008. Toward a Detector-Based Universal Phone Recognizer. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4261-4264. [doi:10.1109/ICASSP.2008.4518596]

Sluijter, A.M.C., van Heuven, V.J., 1996. Acoustic Correlates of Linguistic Stress and Accent in Dutch and American English. Proc. 4th Int. Conf. on Spoken Language, p.630-633. [doi:10.1109/ICSLP.1996.607440]

Sun, X.J., 2002. Pitch Accent Prediction Using Ensemble Machine Learning. Proc. ICSLP, p.953-956.

Taylor, P., 1994. The rise/fall/connection model of intonation. *Speech Commun.*, **15**(1-2):169-186. [doi:10.1016/0167-6393(94)90050-7]

Taylor, P., 1998. The Tilt Intonation Model. Proc. ICSLP, p.1383-1386.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, Massachusetts.

Zhao, J., Yuan, H., Liu, J., Xia, S., 2011. Automatic Lexical Stress Detection Using Acoustic Features for Computer Assisted Language Learning. Proc. APSIPA ASC, p.247-251.

Appendix: Look-up table

Table A1 The look-up table of the articulatory attributes used in the first layer and their corresponding phonemes

Articulatory attribute	International phonetic alphabet
Anterior	/b/, /d/, /ð/, /f/, /l/, /m/, /n/, /p/, /s/, /t/, /θ/, /v/, /z/, /w/
Approximant	/w/, /j/, /l/, /r/, /ə:/, /ə:'/, /əl/
Back	/ai/, /ɔ/, /ɔ'/, /ʌ/, /ʌ'/, /ə/, /ə'/, /au/, /au'/, /əu/, /əu'/, /ɔi/, , /ɔi'/, /u/, /u'/, /u:/, /u:'/, /g/, /k/
Continuant	/ɔ/, /ɔ'/, /æ/, /æ'/, /ʌ/, /ʌ'/, /ə/, /ə'/, /au/, /au'/, /ai/, /ai'/, /ð/, /e/, /e'/, /ə:/, /ə:'/, /r/, /ei/, /ei'/, /l/, /f/, /i/, /i'/, /i:/, /i:'/, /ɔi/, , /ɔi'/, /əu/, /əu'/, /s/, /ʃ/, /θ/, /u/, /u'/, /u:/, /u:'/, /v/, /w/, /j/, /z/
Coronal	/d/, /l/, /n/, /s/, /t/, /z/, /ʒ/
Dental	/d/, /θ/
Fricative	/dʒ/, /tʃ/, /s/, /ʃ/, /z/, /f/, /θ/, /v/, /ð/, /h/, /ʒ/
Glottal	/h/
High	/tʃ/, /i/, /i:/, /dʒ/, /ʃ/, /u/, /u'/, /u:/, /u:'/, /j/, /ei/, /ei'/, /əu/, /əu'/, /g/, /k/, /ŋ/
Labial	/b/, /f/, /m/, /p/, /v/, /w/
Low	/ɔ/, /ɔ'/, /æ/, /æ'/, /au/, /au'/, /ai/, /ai'/, /ɔi/, /ɔi'/, /ʌ/, /ʌ'/, /ə/, /ə'/, /e/, /e'/, /ɔ:/, /ɔ:'/
Mid	/ʌ/, /ʌ'/, /ə/, /ə'/, /e/, /e'/, /ei/, /ei'/, /əu/, /əu'/
Nasal	/m/, /n/, /ŋ/, /əm/, /ən/
Retroflex	/ə:/, /ə:'/, /r/
Round	/au/, /au'/, /əu/, /əu'/, /u:/, /u:'/, /u/, /u'/, /j/, /ɔi/, /ɔi'/, /r/, /w/, /ɔ:/, /ɔ:'/
Stop	/b/, /d/, /g/, /p/, /t/, /k/
Tense	/ɔ/, /æ/, /au/, /ai/, /ei/, /i:/, /əu/, /ɔi/, /u:/, /tʃ/, /s/, /ʃ/, /f/, /θ/, /p/, /t/, /k/, /h/, /i:'/, /i'/, /e'/, /ei'/, /æ'/, /ɔ'/, /au'/ /ai'/, /ʌ'/, /ə'/, /ɔi/, /əu'/, /u'/, /u:'/, /ə:'/, /ɔ:' /
Velar	/g/, /k/, /ŋ/
Voiced	/ɔ/, /ɔ'/, /æ/, /æ'/, /ʌ/, /ʌ'/, /ə/, /ə'/, /au/, /au'/, /ai/, /ai'/, /b/, /d/, /ð/, /e/, /e'/, /ə:/, /ə:'/, /ei/, /ei'/, /g/, /i/, /i'/, /i:/, /i:'/, /dʒ/, /l/, /əl/, /m/, /n/, /ŋ/, /ɔi/, /ɔi'/, /əu/, /əu'/, /r/, /u/, /u'/, /u:/, /u:'/, /v/, /w/, /j/, /z/, /ʒ/
Vowel	/i:/, /i/, /e/, /ei/, /æ/, /ɔ/, /au/, /ai/, /ʌ/, /ə/, /ɔi/, /əu/, /u/, /u:/, /ə:/, /ɔ:/, /i:'/, /i'/, /e'/, /ei'/, /æ'/, /ɔ'/, /au'/, /ai'/, /ʌ'/, /ə'/, /ɔi'/, /əu'/, /u'/, /u:'/, /ə:'/, /ɔ:'/
Silence	Pauses
Unaccented-vowel	/i:/, /i/, /e/, /ei/, /æ/, /ɔ/, /au/, /ai/, /ʌ/, /ə/, /ɔi/, /əu/, /u/, /u:/, /ə:/, /ɔ:/
Accented-vowel	/i:'/, /i'/, /e'/, /ei'/, /æ'/, /ɔ'/, /au'/, /ai'/, /ʌ'/, /ə'/, /ɔi'/, /əu'/, /u'/, /u:'/, /ə:'/, /ɔ:'/