Unsupervised and Semi-supervised Learning of Tone and Pitch Accent

Gina-Anne Levow
University of Chicago
1100 E. 58th St.
Chicago, IL 60637 USA
levow@cs.uchicago.edu

Abstract

Recognition of tone and intonation is essential for speech recognition and language understanding. However, most approaches to this recognition task have relied upon extensive collections of manually tagged data obtained at substantial time and financial cost. In this paper, we explore two approaches to tone learning with substantially reductions in training data. We employ both unsupervised clustering and semi-supervised learning to recognize pitch accent in English and tones in Mandarin Chinese. In unsupervised Mandarin tone clustering experiments, we achieve 57-87% accuracy on materials ranging from broadcast news to clean lab speech. For English pitch accent in broadcast news materials, results reach 78%. In the semi-supervised framework, we achieve Mandarin tone recognition accuracies ranging from 70% for broadcast news speech to 94% for read speech, outperforming both Support Vector Machines (SVMs) trained on only the labeled data and the 25% most common class assignment level. These results indicate that the intrinsic structure of tone and pitch accent acoustics can be exploited to reduce the need for costly labeled training data for tone learning and recognition.

1 Introduction

Tone and intonation play a crucial role across many languages. However, the use and structure of tone varies widely, ranging from lexical tone which determines word identity to pitch accent signalling information status. Here we consider the recognition of lexical tones in Mandarin Chinese syllables and pitch accent in English.

Although intonation is an integral part of language and is requisite for understanding, recognition of tone and pitch accent remains a challenging problem. The majority of current approaches to tone recognition in Mandarin and other East Asian tone languages integrate tone identification with the general task of speech recognition within a Hidden Markov Model framework. In some cases tone recognition is done only implicitly when a word or syllable is constrained jointly by the segmental acoustics and a higher level language model and the word identity determines tone identity. Other strategies build explicit and distinct models for the syllable final region, the vowel and optionally a final nasal, for each tone.

Recent research has demonstrated the importance of contextual and coarticulatory influences on the surface realization of tones.(Xu, 1997; Shen, 1990) The overall shape of the tone or accent can be substantially modified by the local effects of adjacent tone and intonational elements. Furthermore, broad scale phenomena such as topic and phrase structure can affect pitch height, and pitch shape may be variably affected by the presence of boundary tones. These findings have led to explicit modeling of tonal

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 224–231,
New York, June 2006.

context within the HMM framework. In addition to earlier approaches that employed phrase structure (Fujisaki, 1983), several recent approaches to tone recognition in East Asian languages (Wang and Seneff, 2000; Zhou et al., 2004) have incorporated elements of local and broad range contextual influence on tone. Many of these techniques create explicit context-dependent models of the phone, tone, or accent for each context in which they appear, either using the tone sequence for left or right context or using a simplified high-low contrast, as is natural for integration in a Hidden Markov Model speech recognition framework. In pitch accent recognition, recent work by (Hasegawa-Johnson et al., 2004) has integrated pitch accent and boundary tone recognition with speech recognition using prosodically conditioned models within an HMM framework, improving both speech and prosodic recognition.

Since these approaches are integrated with HMM speech recognition models, standard HMM training procedures which rely upon large labeled training sets are used for tone recognition as well. Other tone and pitch accent recognition approaches using other classification frameworks such as support vector machines (Thubthong and Kijsirikul, 2001) and decision trees with boosting and bagging (Sun, 2002) have relied upon large labeled training sets - thousands of instances - for classifier learning. This labelled training data is costly to construct, both in terms of time and money, with estimates for some intonation annotation tasks reaching tens of times realtime. This annotation bottleneck as well as a theoretical interest in the learning of tone motivates the use of unsupervised or semi-supervised approaches to tone recognition whereby the reliance on this often scarce resource can be reduced.

Little research has been done in the application of unsupervised and semi-supervised techniques for tone and pitch accent recognition. Some preliminary work by (Gauthier et al., 2005) employs self-organizing maps and measures of f0 velocity for tone learning. In this paper we explore the use of spectral and standard k-means clustering for unsupervised acquisition of tone, and the framework of manifold regularization for semi-supervised tone learning. We find that in clean read speech, unsupervised techniques can identify the underlying Mandarin tone categories with high accuracy, while even on noisier broadcast news speech, Mandarin tones can be recognized well above chance levels, with English pitch accent recognition at near the levels achieved with fully supervised Support Vector Machine (SVM) classifiers. Likewise in the semi-supervised framework, tone classification outperforms both most common class assignment and a comparable SVM trained on only the same small set of labeled instances, without recourse to the unlabeled instances.

The remainder of paper is organized as follows. Section 2 describes the data sets on which English pitch accent and Mandarin tone learning are performed and the feature extraction process. Section 3 describes the unsupervised and semi-supervised techniques employed. Sections 4 and 5 describe the experiments and results in unsupervised and semi-supervised frameworks respectively. Section 6 presents conclusions and future work.

2 Data Sets

We consider two corpora: one in English for pitch accent recognition and two in Mandarin for tone recognition. We introduce each briefly below.

2.1 English Corpus

We employ a subset of the Boston Radio News Corpus (Ostendorf et al., 1995), read by female speaker F2B, comprising 40 minutes of news material. The corpus includes pitch accent, phrase and boundary tone annotation in the ToBI framework (Silverman et al., 1992) aligned with manual transcription and syllabification of the materials. Following earlier research (Ostendorf and Ross, 1997; Sun, 2002), we collapse the ToBI pitch accent labels to four classes: unaccented, high, low, and downstepped high for experimentation.

2.2 Mandarin Chinese Tone Data

Mandarin Chinese is a language with lexical tone in which each syllable carries a tone and the meaning of the syllable is jointly determined by the tone and segmental information. Mandarin Chinese has four canonical lexical tones, typically described as follows: 1) high level, 2) mid-rising, 3) low falling-rising, and 4) high falling.[1] The canonical pitch con-

[1] For the experiments in this paper, we exclude the neutral tone, which appears on unstressed syllables, because the clear

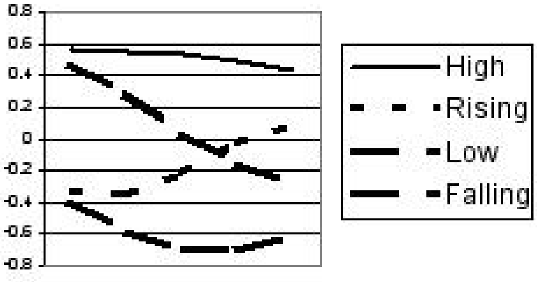

Figure 1: Contours for canonical Mandarin tones

tours for these tones appear in Figure 1.

We employ data from two distinct sources in the experiments reported here.

2.2.1 Read Speech

The first data set is very clean speech data drawn from a collection of read speech collected under laboratory conditions by (Xu, 1999). In these materials, speakers read a set of short sentences where syllable tone and position of focus were varied to assess the effects of focus position on tone realization. Focus here corresponds to narrow focus, where speakers were asked to emphasize a particular word or syllable. Tones on focussed syllables were found to conform closely to the canonical shapes described above, and in previous supervised experiments using a linear support vector machine classifier trained on focused syllables, accuracy approached 99%. For these materials, pitch tracks were manually aligned to the syllable and automatically smoothed and time-normalized by the original researcher, resulting in 20 pitch values for each syllable.

2.2.2 Broadcast News Speech

The second data set is drawn from the Voice of America Mandarin broadcast news, distributed by the Linguistic Data Consortium[2], as part of the Topic Detection and Tracking (TDT-2) evaluation. Using the corresponding anchor scripts, automatically word-segmented, as gold standard transcription, audio from the news stories was force-aligned to the text transcripts. The forced alignment employed the language porting functionality of the University of Colorado Sonic speech recognizer (Pellom et al., 2001). A mapping from the transcriptions to English phone sequences supported by Sonic was created using a Chinese character-pinyin pronunciation dictionary and a manually constructed mapping from pinyin sequences to the closest corresponding English phone sequences.[3]

2.3 Acoustic Features

Using Praat’s (Boersma, 2001) ”To pitch” and ”To intensity” functions and the alignments generated above, we extract acoustic features for the prosodic region of interest. This region corresponds to the “final” region of each syllable in Chinese, including the vowel and any following nasal, and to the syllable nucleus in English.[4] For all pitch and intensity features in both datasets, we compute per-speaker z-score normalized log-scaled values. We extract pitch values from points across valid pitch tracked regions in the syllable. We also compute mean pitch across the syllable. Recent phonetic research (Xu, 1997; Shih and Kochanski, 2000) has identified significant effects of carryover coarticulation from preceding adjacent syllable tones. To minimize these effects consistent with the pitch target approximation model (Xu et al., 1999), we compute slope features based on the second half of this final region, where this model predicts that the underlying pitch height and slope targets of the syllable will be most accurately approached. We further log-scale and normalize slope values to compensate for greater speeds of pitch fall than pitch rise(Xu and Sun, 2002).

We consider two types of contextualized features as well, to model and compensate for coarticulatory effects from neighboring syllables. The first set of features, referred to as ”extended features”, includes the maximum and mean pitch from adjacent syllables as well as the nearest pitch point or points from the preceding and following syllables. These features extend the modeled tone beyond the strict bounds of the syllable segmentation. A second set of contextual features, termed ”difference features”, captures the change in pitch maximum, mean, midpoint, and slope as well as intensity maximum be-

speech data described below contains no such instances.

[2] http://www.ldc.upenn.edu

[3] All tone transformations due to third tone sandhi are applied to create the label set.

[4] We restrict our experiments to syllables with at least 50 ms of tracked pitch in this final region.

tween the current syllable and the previous or following syllable.

In prior supervised experiments using support vector machines(Levow, 2005), variants of this representation achieved competitive recognition levels for both tone and pitch accent recognition. Since many of the experiments for Mandarin Chinese tone recognition deal with clean, careful lab speech, we anticipate little coarticulatory influence, and use a simple pitch-only context-free representation for our primary Mandarin tone recognition experiments. For primary experiments in pitch accent recognition, we employ a high-performing contextualized representation in (Levow, 2005), using both ”extended” and ”difference” features computed only on the preceding syllable. We will also report some contrastive experimental results varying the amount of contextual information.

3 Unsupervised and Semi-supervised Learning

The bottleneck of time and monetary cost associated with manual annotation has generated significant interest in the development of techniques for machine learning and classification that reduce the amount of annotated data required for training. Likewise, learning from unlabeled data aligns with the perspective of language acquisition, as child learners must identify these linguistic categories without explicit instruction by observation of natural language interaction. Of particular interest are techniques in unsupervised and semi-supervised learning where the structure of unlabeled examples may be exploited. Here we consider both unsupervised techniques with no labeled training data and semi-supervised approaches where unlabeled training data is used in conjunction with small amounts of labeled data.

A wide variety of unsupervised clustering techniques have been proposed. In addition to classic clustering techniques such as k-means, recent work has shown good results for many forms of spectral clustering including those by (Shi and Malik, 2000; Belkin and Niyogi, 2002; Fischer and Poland, 2004). In the unsupervised experiments reported here, we employ asymmetric k-lines clustering by (Fischer and Poland, 2004) using code available at the authors’ site, as our primary unsupervised learning approach. Asymmetric clustering is distinguished from other techniques by the construction and use of context-dependent kernel radii. Rather than assuming that all clusters are uniform and spherical, this approach enhances clustering effectiveness when clusters may not be spherical and may vary in size and shape. We will see that this flexibility yields a good match to the structure of Mandarin tone data where both shape and size of clusters vary across tones. In additional contrastive experiments reported below, we also compare k-means clustering, symmetric k-lines clustering (Fischer and Poland, 2004), and Laplacian Eigenmaps (Belkin and Niyogi, 2002) with k-lines clustering. The spectral techniques all perform spectral decomposition on some representation of the affinity or adjacency graph.

For semi-supervised learning, we employ learners in the Manifold Regularization framework developed by (Belkin et al., 2004). This work postulates an underlying intrinsic distribution on a low dimensional manifold for data with an observed, ambient distribution that may be in a higher dimensional space. It further aims to preserve locality in that elements that are neighbors in the ambient space should remain ”close” in the intrinsic space. A semi-supervised classification algorithm, termed ”Laplacian Support Vector Machines”, allows training and classification based on both labeled and unlabeled training examples.

We contrast results under both unsupervised and semi-supervised learning with most common class assignment and previous results employing fully supervised approaches, such as SVMs.

4 Unsupervised Clustering Experiments

We executed four sets of experiments in unsupervised clustering using the (Fischer and Poland, 2004) asymmetric clustering algorithm.

4.1 Experiment Configuration

In these experiments, we chose increasingly difficult and natural test materials. In the first experiment with the cleanest data, we used only focused syllables from the read Mandarin speech dataset. In the second, we included both in-focus (focused) and pre-focus syllables from the read Mandarin speech dataset.[5] In the third and fourth experiments, we chose subsets of broadcast news report data, from the Voice of America (VOA) in Mandarin and Boston University Radio News corpus in English.

In all experiments on Mandarin data, we performed clustering on a balanced sampling set of tones, with 100 instances from each class[6], yielding a baseline for assignment of a single class to all instances of 25%. We then employed a two-stage repeated clustering process, creating 2 or 3 clusters at each stage.

For experiments on English data, we extracted a set of 1000 instances, sampling pitch accent types according to their frequency in the collection. We performed a single clustering phase with 2 to 16 clusters, reporting results at different numbers of clusters.

For evaluation, we report accuracy based on assigning the most frequent class label in each cluster to all members of the cluster.

4.2 Experimental Results

We find that in all cases, accuracy based on the asymmetric clustering is significantly better than most common class assignment and in some cases approaches labelled classification accuracy. Unsurprisingly, the best results, in absolute terms, are achieved on the clean focused syllables, reaching 87% accuracy. For combined in-focus and pre-focus syllables, this rate drops to 77%. These rates contrast with 99-93% accuracies in supervised classification using linear SVM classifiers with several thousand labelled training examples(Surendran et al., 2005).

On broadcast news audio, accuracy for Mandarin reaches 57%, still much better than the 25% level, though below a 72% accuracy achieved using supervised linear SVMs with 600 labeled training examples. Interestingly, for English pitch accent recognition, accuracy reaches 78.4%, aproaching the 80.1% accuracy achieved with SVMs on a comparable data representation.

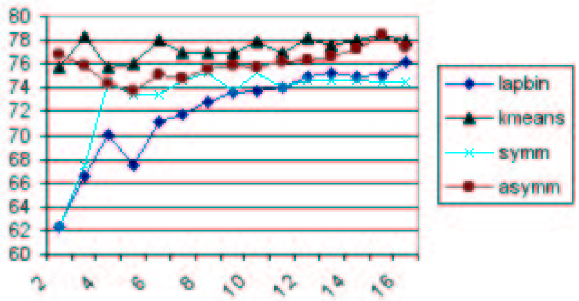

Figure 2: Differences for alternative unsupervised learners across numbers of clusters.

4.3 Contrastive Experiments

We further contrast the use of different unsupervised learners, comparing the three spectral techniques and k-means with Euclidean distance. All contrasts are presented for English pitch accent classification, ranging over different numbers of clusters, with the best parameter setting of neighborhood size. The results are illustrated in Figure 2. K-means and the asymmetric clustering technique are presented for the clean focal Mandarin speech under the standard two stage clustering, in Table 1.

The asymmetric k-lines clustering approach consistently outperforms the corresponding symmetric clustering learner, as well as Laplacian Eigenmaps with binary weights for pitch accent classification. Somewhat surprisingly, k-means clustering outperforms all of the other approaches when producing 3-14 clusters. Accuracy for the optimal choice of clusters and parameters is comparable for asymmetric k-lines clustering and k-means, and somewhat better than all other techniques considered. The careful feature selection process for tone and pitch accent modeling may reduce the difference between the spectral and k-means approaches. In contrast, for the four tone classification task in Mandarin using two stage clustering with 2 or 3 initial clusters, the best clustering using asymmetric k-lines strongly outperforms k-means.

We also performed a contrastive experiment in pitch accent recognition in which we excluded contextual information from both types of contextual features. We find little difference for the majority of

[5] Post-focus syllables typically have decreased pitch height and range, resulting in particularly poor recognition accuracy. We chose not to concentrate on this specific tone modeling problem here.

[6] Sample sizes were bounded to support rapid repeated experimentation and for consistency with the relatively small VOA data set.

	Asymm.	K-means
Clear speech	87%	74.75%

Table 1: Clustering effectiveness for asymmetric k-lines and k-means on clear focused speech.

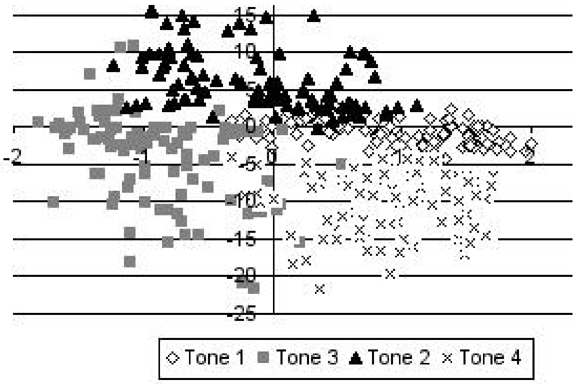

Figure 3: Scatterplot of pitch height vs pitch slope. Open Diamond: High tone (1), Filled black traingle: Rising tone (2), Filled grey square: Low tone (3), X: Falling tone (4)

the unsupervised clustering algorithms, with results from symmetric, asymmetric and k-means clustering differing by less than 1% in absolute accuracy. It is, however, worth noting that exclusion of these features from experiments using supervised learning led to a 4% absolute reduction in accuracy.

4.4 Discussion

An examination of both the clusters formed and the structure of the data provides insight into the effectiveness of this process. Figure 3 displays 2 dimensions of the Mandarin four-tone data from the focused read speech, where normalized pitch mean is on the x-axis and slope is on the y-axis. The separation of classes and their structure is clear. One observes that rising tone (tone 2) lies above the x-axis, while high-level (tone 1) lies along the x-axis. Low (tone 3) and falling (tone 4) tones lie mostly below the x-axis as they generally have falling slope. Low tone (3) appears to the left of falling tone (4) in the figure, corresponding to differences in mean pitch.

In clustering experiments, an initial 2- or 3-way split separates falling from rising or level tones based on pitch slope. The second stage of clustering splits either by slope (tones 1,2, some 3) or by pitch height (tones 3,4). These clusters capture the natural structure of the data where tones are characterized by pitch height and slope targets.

5 Semi-supervised Learning

By exploiting a semi-supervised approach, we hope to enhance classification accuracy over that achievable by unsupervised methods alone by incorporating small amounts of labeled data while exploiting the structure of the unlabeled examples.

5.1 Experiment Configuration

We again conduct contrastive experiments using both the clean focused read speech and the more challenging broadcast news data. In each Mandarin case, for each class, we use only a small set (40) of labeled training instances in conjunction with an additional sixty unlabeled instances, testing on 40 instances. For English pitch accent, we restricted the task to the binary classification of syllables as accented or unaccented. For the one thousand samples we proportionally labeled 200 unaccented examples and 100 accented examples. [7]

We configure the Laplacian SVM classification with binary neighborhood weights, radial basis function kernel, and cosine distance measure typically with 6 nearest neighbors. Following (C-C.Cheng and Lin, 2001), for n-class classification we train $\frac{n(n-1)}{2}$ binary classifiers. We then classify each test instance using all of the classifiers and assign the most frequent prediction, with ties broken randomly. We contrast these results both with conventional SVM classification with a radial basis function kernel excluding the unlabeled training examples and with most common class assignment, which gives a 25% baseline.

5.2 Experimental Results

For the Mandarin focused read syllables, we achieve 94% accuracy on the four-way classification task.

[7] The framework is transductive; the test samples are a subset of the unlabeled training examples.

For the noisier broadcast news data, the accuracy is 70% for the comparable task. These results all substantially outperform the 25% most common class assignment level. The semi-supervised classifier also reliably outperforms an SVM classifier with an RBF kernel trained on the same labeled training instances. This baseline SVM classifier with a very small training set achieves 81% accuracy on clean read speech, but only ≈35% on the broadcast news speech. Finally, for English pitch accent recognition in broadcast news data, the classifier achieves 81.5%, relative to 84% accuracy in the fully supervised case.

6 Conclusion & Future Work

We have demonstrated the effectiveness of both unsupervised and semi-supervised techniques for recognition of Mandarin Chinese syllable tones and English pitch accents using acoustic features alone to capture pitch target height and slope. Although outperformed by fully supervised classification techniques using much larger samples of labelled training data, these unsupervised and semi-supervised techniques perform well above most common class assignment, in the best cases approaching 90% of supervised levels, and, where comparable, well above a good discriminative classifier trained on a comparably small set of labelled data. Unsupervised techniques achieve accuracies of 87% on the cleanest read speech, reaching 57% on data from a standard Mandarin broadcast news corpus, and over 78% on pitch accent classification for English broadcast news. Semi-supervised classification in the Mandarin four-class classification task reaches 94% accuracy on read speech, 70% on broadcast news data, improving dramatically over both the simple baseline of 25% and a standard SVM with an RBF kernel trained only on the labeled examples.

Future work will consider a broader range of tone and intonation classification, including the richer tone set of Cantonese as well as Bantu family tone languages, where annotated data truly is very rare. We also hope to integrate a richer contextual representation of tone and intonation consistent with phonetic theory within this unsupervised and semi-supervised learning framework. We will further explore improvements in classification accuracy based on increases in labeled and unlabeled training examples.

Acknowledgements

We would like to thank Yi Xu for granting access to the read speech data, Vikas Sindhwani, Mikhail Belkin, and Partha Niyogi for their implementation of Laplacian SVM, and Igor Fischer and J. Poland for their implementation of asymmetric clustering.

References

Mikhail Belkin and Partha Niyogi. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceeding of NIPS'02*.

M. Belkin, P. Niyogi, and V. Sindhwani. 2004. Manifold regularization: a geometric framework for learning from examples. Technical Report TR-2004-06, University of Chicago Computer Science.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

C-C.Cheng and C-J. Lin. 2001. LIBSVM:a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

I. Fischer and J. Poland. 2004. New methods for spectral clustering. Technical Report ISDIA-12-04, IDSIA.

H. Fujisaki. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The Production of Speech*, pages 39–55. Springer-Verlag.

Bruno Gauthier, Rushen Shi, Yi Xu, and Robert Proulx. 2005. Neural-network simulation of tonal categorization based on f0 velocity profiles. *Journal of the Acoustical Society of America*, 117, Pt. 2:2430.

M. Hasegawa-Johnson, Jennifer Cole, Chilin Shih abd Ken Chen, Aaron Cohen, Sandra Chavarria, Heejin Kim, Taejin Yoon, Sarah Borys, and Jeung-Yoon Choi. 2004. Speech recognition models of the interdependence among syntax, prosody, and segmental acoustics. In *HLT/NAACL-2004*.

Gina-Anne Levow. 2005. Context in multi-lingual tone and pitch accent prediction. In *Proc. of Interspeech 2005 (to appear)*.

M. Ostendorf and K. Ross. 1997. A multi-level model for recognition of intonation labels. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing Prosody*, pages 291–308.

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

B. Pellom, W. Ward, J. Hansen, K. Hacioglu, J. Zhang, X. Yu, and S. Pradhan. 2001. University of Colorado dialog systems for travel and navigation.

Xiao-Nan Shen. 1990. Tonal co-articulation in Mandarin. *Journal of Phonetics*, 18:281–295.

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8).

C. Shih and G. P. Kochanski. 2000. Chinese tone modeling with stem-ml. In *Proceedings of the International Conference on Spoken Language Processing, Volume 2*, pages 67–70.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labelling English prosody. In *Proceedings of ICSLP*, pages 867–870.

Xuejing Sun. 2002. Pitch accent prediction using ensemble machine learning. In *Proceedings of ICSLP-2002*.

D. Surendran, Gina-Anne Levow, and Yi Xu. 2005. Tone recognition in Mandarin using focus. In *Proc. of Interspeech 2005 (to appear)*.

Nuttakorn Thubthong and Boonserm Kijsirikul. 2001. Support vector machines for Thai phoneme recognition. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(6):803–813.

C. Wang and S. Seneff. 2000. Improved tone recognition by normalizing for coarticulation and intonation effects. In *Proceedings of 6th International Conference on Spoken Language Processing*.

Yi Xu and X. Sun. 2002. Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America*, 111.

C.X. Xu, Y. Xu, and L.-S. Luo. 1999. A pitch target approximation model for f0 contours in Mandarin. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 2359–2362.

Yi Xu. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25:62–83.

Y. Xu. 1999. Effects of tone and focus on the formation and alignment of f0 contours - evidence from Mandarin. *Journal of Phonetics*, 27.

J. L. Zhou, Ye Tian, Yu Shi, Chao Huang, and Eric Chang. 2004. Tone articulation modeling for Mandarin spontaneous speech recognition. In *Proceedings of ICASSP 2004*.