INTERSPEECH 2009 BRIGHTON

Assessing Context and Learning for isiZulu Tone Recognition

Gina-Anne Levow

Department of Computer Science, University of Chicago, Chicago, IL USA

levow@cs.uchicago.edu

Abstract

Prosody plays an integral role in spoken language understanding. In isiZulu, a Nguni family language with lexical tone, prosodic information determines word meaning. We assess the impact of models of tone and coarticulation for tone recognition. We demonstrate the importance of modeling prosodic context to improve tone recognition. We employ this less commonly studied language to assess models of tone developed for English and Mandarin, finding common threads in coarticulatory modeling. We also demonstrate the effectiveness of semi-supervised and unsupervised tone recognition techniques for this less-resourced language, with weakly supervised approaches rivaling supervised techniques.

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 signaling information status.

The majority of research on automatic tone and pitch accent recognition has focused on the East Asian tone languages, in the case of lexical tone, and English or Japanese, for pitch accent. However, by many estimates more than 50% of the world's languages are tonal and come from other tone language families. In this paper, we consider automatic tone recognition for isiZulu, a Bantu language of the Nguni family, with approximately 10 million speakers, most of whom live in South Africa. While there is a rich linguistic literature on the phonology and morphology of Bantu languages [1, 2], there has been little computationally-oriented work in this area. In spite of the large number of speakers, there have been few computational resources or corpora for these languages. Beyond work on tone transcription for Dschang by Bird [3] and recent work on analysis and synthesis of isiZulu and isiXhosa [4, 5], this work constitutes one of the few computational efforts in automatic tone recognition for this class of languages.

As such, study of isiZulu permits us to assess the cross-lingual applicability of approaches and models for tone that have largely been developed and validated on East Asian tone or pitch accent languages. For example, recent research has demonstrated the importance of contextual and coarticulatory influences on the surface realization of tones.[6, 7] The overall shape of the tone or accent can be substantially modified by the local effects of adjacent tone elements, as fundamental physical constraints, such as maximum speed of pitch change, limit the possible tonal transitions. The pitch target approximation model [6] has been proposed to describe and predict the resulting contours. This model has been applied effectively to English and Mandarin and postulates a tonal target consisting of a pitch height and pitch slope target. The approach further argues for an exponential approximation of this tonal target under coarticulatory influences. Since tonal coarticulation is based on inherent physical constraints on control of the vocal apparatus, one would expect the basic coarticulatory effects should be present and must be modeled across tonal languages. However, the nature of the tonal target to be captured may not be as consistent. East Asian tone languages are typically analyzed as having tones characterized by pitch height and contour. In contrast, Bantu tone languages are generally described as having underlying (H)igh or (L)ow (and in some cases Mid) tones, with contours arising from the interaction of different tone targets, rather than from an underlying contour tone. Thus, we consider the impact and effectiveness of coarticulatory tone modeling in this framework for recognition of isiZulu tone.

Also, as observed earlier, resources for isiZulu, such as clear speech corpora with manually labeled tone, are quite rare. Thus, we explore the use of machine learning techniques which require less supervised training data. If such approaches show promise, it will be possible to exploit more readily available unlabeled data from a variety of recordings to improve our understanding and recognition capabilities for isiZulu tone. Here we consider unsupervised k-means clustering to support tone recognition without manual creation of supervised training data and semi-supervised techniques to exploit both small amounts of labeled training data and larger, more readily available sets of unlabeled training data.

The remainder of the paper is organized as follows. We begin with a description of the isiZulu corpus. We then describe the basic tone recognition framework, with baseline features and classifiers. Next we describe the modeling of coarticulatory influences and present the results of contrastive experiments. Later, we describe experiments on semi-supervised and even unsupervised tone recognition for isiZulu, demonstrating promising effectiveness. Finally, we present some summary discussion, conclusions, and future work.

2. Data

In these experiments, we employ a corpus of isiZulu data described in more detail in [4]. The corpus includes 150 utterances, selected from a corpus of sentences from the Web based on bigram graphemic variability to capture phonetic variation. These utterances were read by a native speaker and then manually transcribed, syllabified, and aligned to the audio files. Another native speaker annotated tone based only on sentence text[1]. Tones are aligned with syllables and labeled as either high (H) or low (L)[2]. This alignment yields almost 3000 time-aligned syllables for experimentation. The dominant tone class is the low tone, accounting for approximately 61% of the instances.

[1]The primary goal of Govender *et al.*'s corpus collection was speech synthesis.

[2]We exclude those syllables tagged as 'E', which represent English source words.

3. Tone Modeling

Our model is inspired by the pitch target approximation model of [6]. This approach is grounded in articulatory constraints such as maximum speed of pitch change that predict tonal coarticulation. Each tonal element is viewed as having an underlying target characterized by pitch slope and height. Under coarticulatory constraints, the target may not be achieved immediately, but is gradually approached, with the difference decaying exponentially.

In addition to earlier approaches that employed phrase structure [8], several recent approaches to tone recognition in East Asian languages [9, 10, 11] and to tone generation [12] 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. With StemML[12], templates corresponding to canonical tone models are presumed to be deformed to conform to the current context. Studies of pitch accent have often included features providing contrasts with neighboring words or syllables, though less explicitly in a coarticulatory framework [13]. [14]'s work captures elements of local influence on accent identity, applying the pitch target approximation model to English pitch accent recognition.

Here, we take the syllable as the domain of tone prediction, consistent with [14]. We employ an acoustic model at the syllable level, employing pitch, intensity and duration measures. In contrastive experiments, we also exploit word boundary information. The acoustic measures are computed using Praat's [15] "To pitch" and "To intensity" functions. We compute log-scaled and speaker-normalized forms for all pitch and intensity values.

We compute two sets of features: one set describing features local to the syllable and one set capturing contextual information.

3.1. Local features

We extract features to represent the pitch height and pitch contour of the syllable, consistent with the components of the pitch target approximation model. For pitch features, we extract the following information:

- pitch values for five evenly spaced points in the voiced region of the syllable

  We perform piecewise cubic interpolation of missing values.
- pitch maximum, mean, minimum, and range
- pitch slope

  Following [16], we assume that the pitch target can be expected to be closely approached by the middle of the syllable. Thus, we compute a linear fit to pitch slope from the midpoint to the end of the syllable.

We also obtain the following non-pitch features:

- intensity maximum and mean
- syllable duration
- syllable position

  To capture effects such as pitch reset and downdrift associated with phrase initiation and position, we compute syllable position from the beginning and end of each pseudo-phrase, identified as a silence delimited interval.

3.2. Context Modeling

To capture local contextual influences and cues, we explore both the addition of new features and the modification of base features. First, we consider the addition of two sets of features. The first set of features ("difference features") corresponds to differences between the current syllable and its preceding and following syllables. They include difference between pitch maxima, pitch means, pitches at the midpoint of the syllables, pitch slopes, intensity maxima, and intensity means. The second set of features, which we will refer to as "extended syllable" features, are simply the last pitch values from the end of the preceding syllable and the first from the beginning of the following syllable.

These features are intended to capture both the relative differences in pitch associated with tone as well as to compensate for phenomena such as pitch peak delay in which the actual target of a high or rising tone may not be reached until the following syllable.

Finally, we consider the word context in which the syllable appears. The morphological structure of the word determines its surface tonal realization, and prior research has indicated that tonal patterns and syllable strength within the domain of the word also affect tonal phonology and phonetics. In this case, we incorporate word information by adding two features:

- the difference between the mean pitch of the current word and the mean pitch of the current syllable, and
- the difference between the mean intensity of the current word and the mean intensity of the current syllable.

We also replace each of the five original pitch point values with the corresponding difference between the original value and the mean pitch of the word.

4. Supervised Classifier

For all supervised experiments reported in this paper, we employ a Support Vector machine (SVM) with a linear kernel. Support Vector Machines provide a fast, easily trainable classification framework that has proven effective in a wide range of application tasks. For example, in the binary classification case, given a set of training examples presented as feature vectors of length D, the linear SVM algorithm learns a vector of weights of length D which is a linear combination of a subset of the input vectors and performs classification based on the function $f(x) = sign(w^T x - b)$. Furthermore, SVMs have been generalized from binary classification to multiclass classification as well as semi-supervised frameworks. The corresponding weights can also provide insight into the contribution of different features to the classification process. We employ the publicly available multi-class implementation of SVMs, LIBSVM [17]. We use four-fifths of the data for training and one-fifth for testing.

5. Results for Context Modeling

We assess the effects of different contextual features for tone modeling. Our context modeling experiments consider two primary contrastive conditions: context encoding - "extended" or "difference" features- and context position - preceding, following, both, or none. The results for these contrasts using the Support Vector Machine classifier appear in Table 1. Clearly, con-

	Extended	Difference	Both
Position			
None	74.1%	74.1%	74.1%
Following	74.6%	74.6%	74.8%
Preceding	75.3%	76.5%	76.5%
Both	74.7%	76%	76.2%

Table 1: Tone Classification Varying Context Encoding and Position

textual evidence improves over the no-context case. In particular, modeling preceding context yields greater improvements than modeling following context alone. Furthermore, for this collection, the use of "difference" features produces little difference in effectiveness from the use of "extended" features. Most common class assignment would yield 61% accuracy.

The greater importance of preceding context over following context is consistent with analysis of coarticulation that argues for a greater role of carryover co-articulation from preceding syllables than anticipatory co-articulation with following syllables [18]. This finding is also consistent with classification results for English pitch accent and Mandarin Chinese tone [19].

Finally, we consider the impact of the word-based feature set on tone recognition accuracy. Here, the full context model reaches an accuracy of 76%. Interestingly, the 'no context' model also achieves an accuracy of 76%. The word-based normalization concisely captures the contextual influences on tone.

6. Unsupervised isiZulu Tone Recognition

Since only relatively small amounts of tone-labeled isiZulu data are available, we explore the use of minimally supervised techniques to identify tone categories. In particular, we employ unsupervised clustering to distinguish High and Low tones. There has been significant recent interest unsupervised clustering, not only with standard k-means approaches, but also with a range of spectral clustering techniques, that cluster based on a spectral decomposition of a neighborhood or affinity matrix [20, 21, 22]. Interestingly, prior experiments on unsupervised tone and pitch accent clustering in Mandarin Chinese and English [23], respectively, found that for these tasks, k-means clustering performed very competitively with the more computationally demanding spectral clustering approaches. Thus, here we will employ k-means clustering as our primary condition.

6.1. Clustering Experiments

We perform experiments on the same sample set from the corpus as used in the supervised experiments, with 61% of syllables bearing low tone and 39% bearing high tone. We use all the samples in the clustering process and use the same subset as the test set. We employ the best feature set from the supervised experiments. We compare different numbers of clusters and evaluate the clustering by assigning the most frequent label in each cluster to all members of the cluster.

The results appear in Table 2. We create between two and six clusters using k-means clustering. The best clustering is achieved with three clusters and an accuracy of 75.3%, approaching supervised levels. All clusters are well above the baseline chance effectiveness of 61%. This success indicates that the isiZulu tones are well-separated in acoustic space.

# Clusters	2	3	4	5
Accuracy	71%	75.3%	73.8%	73.1%

Table 2: Unsupervised clustering of isiZulu tone is competitive with supervised approaches.

# labeled	1000	500	100
Accuracy			
Semi-supervised	78.5%	76.3%	73.3%
Supervised	76%	71%	72.8%

Table 3: Comparison of semi-supervised and supervised learning of isiZulu tone

7. Semi-supervised Learning of isiZulu Tone

Semi-supervised machine learning approaches aim to exploit the information in more readily available unlabeled data, in conjunction with the evidence from a smaller amount of labeled training materials. Given the effectiveness of the clustering approaches above, it seems likely that semi-supervised techniques which exploit the structure of the unlabeled data should be effective.

Rather than employing multiple learners in co-training or more direct self-training for semi-supervised learning, we employ learners in the Manifold Regularization framework developed by [24]. 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, an extension of Support Vector Machines termed "Laplacian Support Vector Machines", allows training and classification based on both labeled and unlabeled training examples.

7.1. Semi-supervised Experiments

We continue to employ the same test set used in all prior experiments, again with the best feature set. We compare tone recognition accuracy with different amounts of training data: 1000, 500, and 100 labeled examples, with the remainder of the samples used as unlabeled data in a transductive setting. We configure the Laplacian SVM classification with binary neighborhood weights, radial basis function kernel, and cosine distance measure, and 3 nearest neighbors. We then contrast the effectiveness of the semi-supervised classifier with that of the supervised SVM setting.

Results appear in Table 7.1. The semi-supervised approach outperforms the supervised approach with comparable amounts of data. Good accuracy is maintained as the amount of labeled data is reduced. For less-resourced languages such as isiZulu, it is particularly useful that effective tone recognition can be performed even with little or no training data.

8. Discussion, Conclusion, and Future Work

We have assessed contextual modeling, unsupervised clustering, and semi-supervised learning for tone recognition in isiZulu, a language of the Bantu family. We find that modeling of tonal coarticulation with contextual features improves

tone recognition accuracy. In particular, compensating for carryover coarticulation through modeling the preceding syllabic context yields the greatest improvement. Further, we find that word-based normalization provides comparable compensation for contextual and coarticulatory influences. Finally, unsupervised and semi-supervised approaches to tone classification show promise for working with this less-resourced language by enabling categorization competitive with fully supervised techniques.

These findings are largely consistent with prior work on prosodic labelling for East Asian tone and pitch accent languages, such as Mandarin Chinese and English, respectively. We find consistent effects for context modeling, as we expected based on the common basis of coarticulatory and contextual constraints across tone types. Likewise, the tonal categories in isiZulu are sufficiently well-separated in the acoustic space that unsupervised and semi-supervised techniques that exploit this structure yield good effectiveness for this language as well.

Future research will investigate the integration of additional feature types, such as band energy and voice quality. Duanmu [25] states that in many tone languages, tones in low register are associated with particular voice quality, citing work on isiZulu. Such measures have proven useful in recognizing pitch accent in English [26] and neutral tone in Mandarin Chinese [27]. We will also pursue the use of other unsupervised and semi-supervised approaches to support improvements in tone recognition for this less-resourced language.

9. Acknowledgments

We would like to thank Etienne Barnard and Natasha Govender for access to the isiZulu data set and annotations, as well as C-C.Cheng and C-J. Lin for the implementation of LibSVM. This work was supported by NSF IIS: 0414919.

10. References

[1] G. N. Clements and J. Goldsmith, *Autosegmental studies in Bantu tone*. Foris Publication, 1984.

[2] G. Poulos and C. T. Msimang, *A Linguistic Anaylsis of Zulu*. Via Afrika, 1998.

[3] S. Bird, "Automated tone transcription," in *Proceedings of Workshop on Computational Phonology*, 1994, pp. 1–12.

[4] N. Govender, E. Barnard, and M. Davel, "Fundamental frequency and tone in isizulu: initial experiments," in *INTERSPEECH-2005*, 2005, pp. 1417–1420.

[5] ——, "Pitch modelling for the Nguni languages," *South African Computer Journal*, vol. 38, pp. 28–39, 2007.

[6] Y. Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol. 25, pp. 62–83, 1997.

[7] X.-N. Shen, "Tonal co-articulation in Mandarin," *Journal of Phonetics*, vol. 18, pp. 281–295, 1990.

[8] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*. Springer-Verlag, 1983, pp. 39–55.

[9] C. Wang and S. Seneff, "Improved tone recognition by normalizing for coarticulation and intonation effects," in *Proceedings of 6th International Conference on Spoken Language Processing*, 2000.

[10] J. L. Zhou, Y. Tian, Y. Shi, C. Huang, and E. Chang, "Tone articulation modeling for Mandarin spontaneous speech recognition," in *Proceedings of ICASSP 2004*, 2004.

[11] N. Thubthong and B. Kijsirikul, "Support vector machines for Thai phoneme recognition," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 6, pp. 803–813, 2001.

[12] C. Shih and G. P. Kochanski, "Chinese tone modeling with stem-ml," in *Proceedings of the International Conference on Spoken Language Processing, Volume 2*, 2000, pp. 67–70.

[13] M. Ostendorf and K. Ross, "A multi-level model for recognition of intonation labels," in *Computing Prosody*, Y. Sagisaka, N. Campbell, and N. Higuchi, Eds., 1997, pp. 291–308.

[14] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proceedings of ICSLP-2002*, 2002.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9–10, pp. 341–345, 2001.

[16] X. Sun, "The determination, analysis, and synthesis of fundamental frequency," Ph.D. dissertation, Northwestern University, 2002.

[17] C-C.Cheng and C.-J. Lin, "LIBSVM:a library for support vector machines," 2001, software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[18] Y. Xu, "Production and perception of coarticulated tones," *Journal of Acoustic Society of America*, vol. 95, no. 4, pp. 2240–2253, 1994.

[19] G.-A. Levow, "Context in multi-lingual tone and pitch accent prediction," in *Proceedings of Interspeech 2005*, 2005, pp. 1809–1812.

[20] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, 2000.

[21] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proceeding of NIPS'02*, 2002.

[22] I. Fischer and J. Poland, "New methods for spectral clustering," IDSIA, Tech. Rep. ISDIA-12-04, 2004.

[23] G.-A. Levow, "Unsupervised and semi-supervised tone and pitch accent recognition," in *Proceedings of HLT-NAACL 2006*, 2006, pp. 224–231.

[24] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from examples," University of Chicago Computer Science, Tech. Rep. TR-2004-06, 2004.

[25] S. Duanmu, "Tone: An overview," *Glot International*, vol. 2, no. 4, 1996.

[26] A. Rosenberg and J. Hirschberg, "On the correlation between energy and pitch accent in read english speech," in *INTERSPEECH-2006*, 2006, pp. 1294–1297.

[27] D. Surendran and G.-A. Levow, "Additional cues for mandarin tone recognition," University of Chicago, Computer Science, Tech. Rep. TR-2006-04, 2006.