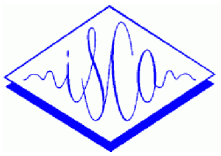
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Nativeness Classification with Suprasegmental Features on the Accent Group Level

Mahnoosh Mehrabani, Joseph Tepperman, and Emily Nava*

Rosetta Stone Labs
Boulder, Colorado, USA

mahmehrabani@utdallas.edu, {jtepperman,enava}@rosettastone.com

10.21437/Interspeech.2012-553

Abstract

We present a novel approach to discriminating native and nonnative utterances based on suprasegmental features extracted at the Accent Group (AG) level. Past studies have shown modeling a set of shared intonation patterns across AGs to be effective in predicting local f0 contour shapes. Here we demonstrate that AG level prosodic features are also effective in nativeness classification. The proposed suprasegmental feature set is very low dimensional, and is derived from f0 and energy contours across the AG, as well as normalized duration of the syllables within each AG. A Random Forest back end classifier is used to combine AG level scores from GMM and Decision Tree models, producing nativeness scores at the utterance level. The proposed prosodic nativeness classifier achieves 83.3% accuracy for 2-AG utterances and 89.1% accuracy for 3-AG utterances, exceeding a baseline Gaussian Supervector system's performance by more than 10% absolute. The vastly lower dimensionality of the proposed feature set relative to the baseline method suggests the importance of suprasegmental features over traditional spectral cues in contributing to the perceived nativeness of a learner's language.

Index Terms: nativeness, prosody, intonation, rhythm

1. Introduction

Mastering the suprasegmental characteristics of English–including rhythm, stress, and intonation–is one of the most widely encountered difficulties among English learners [1]. Nonnative speakers tend to impose prosodic patterns from their first language onto their interpretation and pronunciation of any foreign language, a process known as transfer. Because of the importance of suprasegmental pronunciation to learner perception and intelligibility [1], it is crucial for learners of English to study the suprasegmental patterns even in the early stages of learning. Some studies, such as [2], suggest English rhythmic entrainment through repetition of phrases sharing common, iconic accent patterns. These pedagogical concerns have motivated researchers in automatic prosodic assessment to search for a common unit of analysis that can capture learner pronunciation variability on many suprasegmental levels simultaneously, while allowing for an assessment paradigm that is generalizable to any arbitrary utterance.

The parroting paradigm of second language instruction [3]–i.e. one in which learners repeat phrases after a native-speaker prompt–naturally lends itself to template based scoring algorithms such as those proposed in [4], where explicit comparisons between the learner and the native target are made without recourse to deeper levels of model abstraction. However, the limitations of this approach to scoring should be obvious: students are assessed relative to one particular native realization rather than to native speech in general, and no learner utterance can be scored unless it is first recorded by a native speaker.

A favorable alternative, both for pedagogical and technical reasons, is to conduct analysis at an intermediate linguistic level that is large enough to capture the suprasegmental variability that characterizes native speech, but also sufficiently short and abstract so that such a unit could be recombined into any given sentence. Because it satisfies both the generalizability and variability requirements, in this study we have selected the Accent Group (AG) as our unit of suprasegmental analysis.

The Accent Group [5] is defined as a level in the prosodic hierarchy existing between the Foot and the Intermediate Phrase. For purposes of this study, we use the specification given in [6, 7], where the term Left Headed Foot was suggested to refer to what is in reality an Accent Group: an accented syllable and all of the unaccented syllables following it (A Foot is technically defined in terms of stress, not accent–and hence more than one Foot can fall between any two adjacent accents.) Though their terminology differed, the work in [6, 7] showed that Accent Group based factorization effectively predicts local frequency contour shapes for speech synthesis.

Suprasegmental features have been applied to many areas of automatic speech processing, including speaker, language, dialect, and accent classification [8, 9], and there exist many approaches to modeling suprasegmental features at various levels. Frame level features including statistics of f0, and energy contours have been used for speaker identification [10]. However, the local dynamics of intonation also encode valuable information that long-term statistics do not capture. For example, pitch movement patterns between adjacent frames have been used in dialect distance assessment [11]. Furthermore, larger-scale intonation movements have been modeled by fitting a piecewise linear model to the f0 contour [8]. Recently, pseudosyllable units have become increasingly popular as speech units for suprasegmental feature extraction, especially in text-independent tasks [9, 12]. The work in [13] presents a nativeness classifier for English based on acoustic and prosodic features. Their prosodic features are similar to those presented in [9] for language identification, which are based on an approximation of f0 and energy contours at the pseudosyllable level.

In this work, we show that suprasegmental features derived at the Accent Group level can classify nativeness in a corpus of parroted speech prompts, with accuracy comparable to or better than that of state-of-the-art spectral speech features extracted at

* Mahnoosh Mehrabani is currently with the University of Texas at Dallas Center for Robust Speech Systems.

	Japanese Learners	*Native English Speakers*
Female Speakers	51	5
Male Speakers	44	11
Total Phrases	1478	1592
Total Hours	1.3	1.4

Table 1: Speech corpus statistics.

the more traditional frame level. Eventually such an algorithm can be adapted to generate automatic suprasegmental scores for arbitrary learner utterances in the context of a computer-assisted language learning application.

2. Corpus

The speech data used in the experiments described below is divided into two sub corpora: Japanese learner parroting, and native English speaker parroting. All speakers parroted some subset of 100 English phrases, hand-selected to represent a variety in length, subject matter, and suprasegmental content. Statistics about these sets are given in Table 1. The reference prompts were produced by 4 professional voicers (2 male, 2 female) and were previously used as actual prompts in the Rosetta Stone Version 4 American English product. The learner parroting recordings were collected in Japan, all from native Japanese speakers. Though none of them were fluent in English, the speakers represented a wide range of English proficiency. All recordings selected for these experiments were checked to ensure that they were at least devoid of noise and that they were in-grammar; that is, that the student produced all of the words in the prompt. These learners were not explicitly told to parrot the reference prompts on the suprasegmental level. The native speakers of English were all in-house employees of Rosetta Stone, and none of them were professional product voicers. Each speaker was explicitly instructed to try to match the voicer's rhythm and intonation, and they were allowed to listen to the prompt and record their own version as many times as they wanted. Though the recording conditions were not identical between the native and nonnative populations, this yielded two very different data sets: one with presumably proficient native parroting, and one with learner parroting of diverse proficiency. Automatic speech segmentation was done using forced Viterbi decoding of the target utterance using Rosetta Stone's proprietary speech recognition system. The segmentation process provided both word level and phoneme level alignments of the speech data. The decoded sequence of phonemes was then chunked into syllables based on each word's expected syllabification according to a pronunciation dictionary. The decoding grammar allowed for possible word deletion and silence insertion, as to be expected in learner speech. f0 and energy contours were estimated for each utterance. In addition, the locations of pitch accents in the prompts (and hence the AGs) were annotated by two experts, and any disagreement between them was resolved through closer listening. AGs for the parroting speakers were defined relative to the corresponding prompt.

3. Accent Group level prosodic features

In this study we extract suprasegmental features at Accent Group level. As mentioned, an Accent Group is defined as an accented syllable followed by all unaccented syllables until the next accent or phrase boundary. In this study, we limit our analysis to polysyllabic Accent Groups (an accented syllable followed by at least one unaccented syllable.) We extract three categories of suprasegmental features based on f0, energy, and duration within the Accent Group.

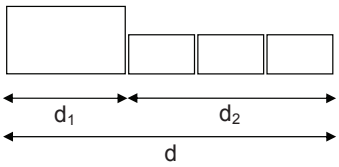


Figure 2: *Duration features: d_1 represents the accented syllable, d_2 the unaccented syllables. Both are normalized by d.*

The intonation features are based on a polynomial approximation of the f0 contour. First, unvoiced segments are linearly interpolated to achieve a continuously voiced estimate. Next, these interpolated f0 contours are modeled with n-degree polynomials:

$$P_1 x^n + P_2 x^{n-1} + \cdots + P_n x + P_{n+1} \quad (1)$$

The coefficients are concatenated to obtain one $(n+1)$ dimensional feature vector of coefficients per Accent Group: $\{P_1, \ldots, P_{n+1}\}$.

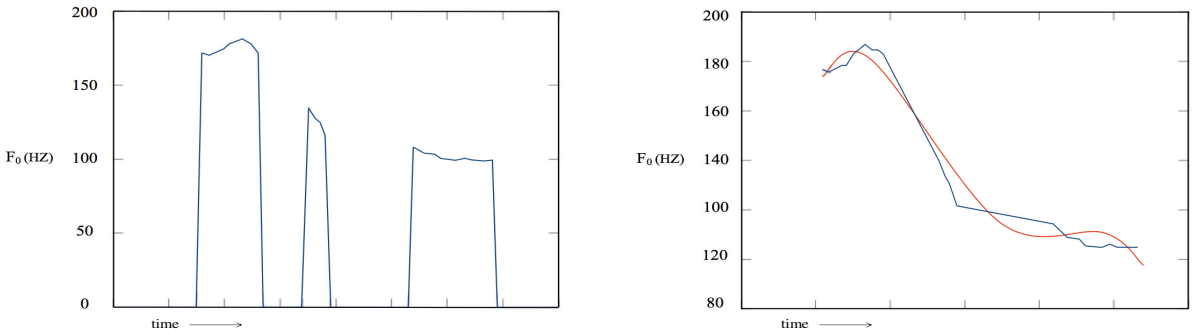
Polynomial curve fitting is performed in a least-squares sense. Here we chose the degree n=5 and reduced the coefficient vectors to 4-dimensional, since the 5th coefficient is a bias scalar representing the average pitch over the contour. Fig. 1 shows an example of an f0 contour polynomial approximation. Fig. 1(a) is the original f0 estimate, and Fig. 1(b) represents the interpolated and approximated versions of the contour. In this case, the polynomial approximation seems to capture the shape of the f0 movement with only 5 coefficients. We perform the same polynomial approximation method for the energy contour, to obtain a 4-dimensional coefficient vector representation of the energy over each Accent Group.

We also estimated duration based features for each Accent Group. Since the length and linguistic content of each Accent Group can vary significantly, the absolute duration in seconds would probably not offer much discriminative power. Instead, we use relative durations of the accented and unaccented syllables, since those are common to all Accent Groups as we have defined them. Fig. 2 illustrates a polysyllabic Accent Group with three unaccented syllables. The big box represents the accented syllable at the beginning of the AG, and the smaller boxes represent the unaccented syllables. The proposed duration vector we estimated consists of three features: the normalized duration of the accented syllable, d_1/d; the normalized duration of the sequence of unaccented syllables, d_2/d; and the total number of syllables in the Accent Group. These three features approximate a very rough encoding of the rhythmic pattern delineated by the syllables in the Accent Group. These intonation, energy, and duration feature vectors were concatenated to obtain one 11-dimensional prosodic feature per Accent Group, which was then used for utterance level classification, as explained below.

4. Classification and results

4.1. Baseline

For comparison with our proposed suprasegmental method, we attempted to duplicate the baseline system reported in [13]. Note that, here we only implement their acoustic classifier as



(a) (b)

Figure 1: *(a) Original f0 estimate. (b) Polynomial approximation of the contour, over the interpolated contour from (a).*

our baseline. Their approach adapted the Gaussian Supervector (GSV) speaker verification technique [14] to the nativeness classification task. A GSV is a very long vector of features created by concatenating the means of a trained Gaussian Mixture Model (GMM). A Universal Background Model (UBM) is first trained on all available training data; this takes the form of an n-mixture GMM. For each test file, one iteration of Maximum a Posteriori (MAP) model adaptation is performed on the UBM, to create a new, unique GMM for each test file. The means of that adapted UBM are then stacked to create the feature vectors for nativeness classification, and then those vectors are classified using an external machine learning algorithm; in the case of [13], that algorithm was a Support Vector Machine (SVM).

We used 39 PLP features (13 plus first and second order delta coefficients) extracted using a 20 msec window at 10 msec intervals, according to HTK's implementation. The UBM was trained with flat start initialization and an iterative procedure of repeated alignment and embedded reestimation. A single-mixture GMM silence model was trained simultaneously, to isolate the nonspeech frames of each audio file. The number of UBM mixtures was set to 64, since that is what achieved the best performance in [13] for speech files of 3–10 seconds in length (which is roughly the length of all utterances in our corpus). As in [13], MAP adaptation was done on only the UBM means, with a relevance factor of 16. The SVM for final nativeness classification was the libSVM implementation, using a linear kernel. In all experiments, training and testing was conducted using a leave-one-out speaker level cross validation.

4.2. Prosodic

The results of nativeness classification based on proposed Accent Group level prosodic features are presented in this section. All presented results are based on cross validation. In order to make sure that nativeness is classified, rather than the speakers being classified, in each iteration we set aside all the utterances from one speaker and trained native and nonnative models. Next, the utterances that were not present in training were tested against the models. Finally, the number of all correctly classified utterances in all the iterations were divided by the whole number of utterances to obtain an average accuracy.

We first present the classification results at the AG level, which means that instead of each utterance, each Accent Group is classified as native or nonnative. This is a more difficult task than classifying utterances. Table 2 shows the classification results. Three different methods were used to model and classify prosodic features: GMM with 128 mixtures, Decision Tree, and SVM with polynomial kernel. For training of SVM and Decision Tree models, the WEKA data mining software [15] was used: SMO for support vector classifier and J48 for decision tree classifier. The results show that GMMs and Decision Trees perform better than SVMs.

	Native	*Nonnative*	*Overall*
GMM	64.5	64.5	64.5
Decision Tree	64.2	64.5	64.4
SVM	68.5	52.1	60.3

Table 2: Classification results (%) on the Accent Group level, before utterance level classification.

Next, utterances were classified using a Random Forest back end classifier to combine Accent Group scores. A Random Forest is a collection of decision trees such that each tree depends on an independently sampled random vector identically distributed for all the trees in the forest [16]. Random Forest is one of the most accurate learning algorithms available, and our experiments show that it achieves better results than other classifiers as a back end to combine scores. WEKA software [15] was used for Random Forest training with 100 trees. Table 3 and 4 show the classification results for 2-AG utterances, and 3-AG utterances, respectively. In both tables, the first row presents the results of combining GMM AG scores, the second row shows the results of combining Decision Tree AG scores, and finally the third row presents a fusion of GMM and Decision Tree AG scores. When fusing AG scores, we incorporated the number of syllables for each Accent Group as additional features to the GMM and/or Decision Tree scores.

5. Discussion

Between Tables 3 and 4, we see an overall improvement in classification accuracy for utterances with 3 AGs (rather than 2). This is not surprising; not only do the proposed suprasegmental classifiers have more features to work with in the 3-AG case,

	Native	*Nonnative*	*Overall*
GMM	71.0	71.4	71.2
Decision Tree	81.2	81.0	81.1
GMM + Decision Tree	83.0	83.6	83.3
GSV baseline	56.0	91.0	72.1

Table 3: Classification results (%) for 2-AG utterances.

	Native	*Nonnative*	*Overall*
GMM	84.8	62.5	75.7
Decision Tree	87.5	83.5	86.8
GMM + Decision Tree	90.6	86.9	89.1
GSV baseline	60.8	89.0	73.8

Table 4: Classification results (%) for 3-AG utterances.

but utterances with more AGs tend to have longer absolute durations and more frames for adaptation on the GSV baseline. This baseline's performance is comparable to that reported in [13], which achieved 66.7% accuracy using a 64-mixture GSV on utterances of similar length to those in our corpus. However, their reported accuracy was much more balanced between the native and nonnative populations, whereas in ours there is a huge disparity in accuracy, with the nonnative performance exceeding that of the native speakers. This can be explained by the unbalanced nature of our corpus: we had an order of magnitude more nonnatives than natives (see Table 1), while the data used in [13] was balanced almost 50/50. With the speaker level cross validation procedure in these experiments, the GSV baseline was adversely affected by these corpus conditions, while the non baseline methods were not. The reason is based on two main differences between the proposed system and the baseline in terms of features and classification method. First, the proposed suprasegmental features on the AG level encode longer term information that is less speaker dependent compared to frame level features in the baseline. Second, modeling native and nonnative AG level features, and then combining scores from obtained models for utterance level classification introduces more training samples compared to baseline, which trains one supervector per utterance.

The most important outcome of Tables 3 and 4 is the superiority of the AG based methods over the baseline in the overall classification results. This is notable especially considering the much smaller set of features used in the AG based cases. For each Accent Group, a total of only 11 features were used to encode. And at most, only three such Accent Groups were used for utterance level classification. Compare this to the baseline's 39 PLP features, and then the 2496 (= 39 means times 64 mixtures) features used in SVM utterance level classification. Clearly there is very valuable nativeness information encoded in these relatively few and simple suprasegmental cues.

6. Conclusion

This paper has demonstrated a new method for suprasegmental analysis of speaker nativeness, one that outperforms a baseline system while using vastly fewer features. Because analysis is conducted on the level of the Accent Group, this method can be applied to any arbitrary sentence in a second language practice curriculum, as long as the Accent Group segmentation is known. One limitation of this work was the way in which the Accent Groups were defined a priori by human annotators, based on the acoustics of the prompts. For maximum flexibility in the authoring of second-language practice content, future research in this area will be well served to automatically segment all utterances (or perhaps just the target prompts) into Accent Groups, either using acoustic information as in [17], or based on the syntax of the text alone, as in [18]. The relationship between these Accent Group features and those derived from pseudosyllables as in [13], requires further examination, as do features based on other prosodic units. Future work could also include applying the proposed nativeness classifier to a wider variety of nonnative English speakers including more L1 languages.

7. References

[1] D. Felps and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 18, pp. 1030–1040, 2010.

[2] W. Stannard-Allen, *Living English Speech*. Longmans, 1954.

[3] M. Tanner and M. Landon, "The effects of computer-assisted pronunciation readings on ESL learners' use of pausing, stress, intonation, and overall comprehensibility," *Language Learning & Technology*, vol. 13, no. 3, pp. 51–65, October 2009.

[4] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental english through parroting," in *Proceedings of Speech Prosody*, 2010.

[5] R. Ogden, S. Hawkins, J. House, M. Huckvale, J. Local, P. Carter, J. Dankovicova, , and S. Heid, "Prosynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis," *Computer Language and Science*, pp. 177–210, 2000.

[6] E. Klabbers, J. Van Santen, and J. Wouters, "Prosodic factors for predicting local pitch shape," in *Proc. 2002 IEEE Workshop on Speech Synthesis*, pp. 123–126.

[7] E. Klabbers and J. Santen, "Clustering of foot-based pitch contours in expressive speech," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[8] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, vol. 7, 1998, pp. 3189–3192.

[9] C.-Y. Lin and H.-C. Wang, "Language identification using pitch contour information," in *Proc. IEEE ICASSP*, 2005, pp. 601–604.

[10] M. Carey, E. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *ICSLP*, vol. 3, 1996, pp. 1800–1803.

[11] M. Mehrabani, H. Boril, and J. Hansen, "Dialect distance assessment method based on comparison of pitch pattern statistical models," in *IEEE ICASSP*, 2010, pp. 5158–5161.

[12] J.-L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1904–1911, 2007.

[13] J. Lopes, I. Trancoso, and A. Abad, "A nativeness classifier for ted talks," in *IEEE ICASSP*, 2011, pp. 5672–5675.

[14] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 15, May 2006.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[16] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] H. Shimodaira and M. Nakai, "Accent phrase segmentation using transition probabilities between pitch pattern templates." in *Proc. EuroSpeech*, September 1993, pp. 1767–1770.

[18] J. Tepperman and E. Nava, "Where should pitch accents and phrase breaks go? a syntax tree transducer solution," in *Proc. Interspeech*, August 2011.